Systematic interaction interface and variant characterization using protein interaction profiling

Dissertation

Zur Erlangung des Grades

Doktor der Naturwissenschaften

Am Fachbereich Biologie

Der Johannes Gutenberg-Universität Mainz

Dalmira Hubrich

geb. am 28.07.1993 in Qostanay, Kazakhstan

Mainz, Oktober 2024

Dekan: Prof. Dr. Eckhard Thines

1. Berichterstatter:

2. Berichterstatter:

Tag der mündlichen Prüfung: 21.10.2024

"The important thing is to never stop questioning." – Albert Einstein"

Acknowledgements

Contents

Chapter 1

Introduction

1.1 The Complexity of Human Genetic Variation

Genetic variation is a primary factor in evolution, driving the appearance of new phenotypes with various degrees of adaptability to environmental factors. A human genetic variation is defined as the diversity in DNA sequences and genetic characteristics among the individuals within populations (**Alberts et al. 2002**).

These variations arise from replication errors or spontaneous nucleotide alterations that occur in DNA replication during cell division. In addition to these endogenous factors, exogenous influences such as radiation or chemicals can also cause changes in the genome. Genetic variation occurs at different scales from structural to single-point mutations. Structural variants are usually found to have a size of 1Mb and happen on chromosome level (e.g.fragile sites) (**Chaisson et al. 2019**). On the contrary, small variants span from duplications, deletions, insertions and inversions to short nucleotide polymorphisms or SNPs (**Nesta et al. 2021**) Figure 1.1).

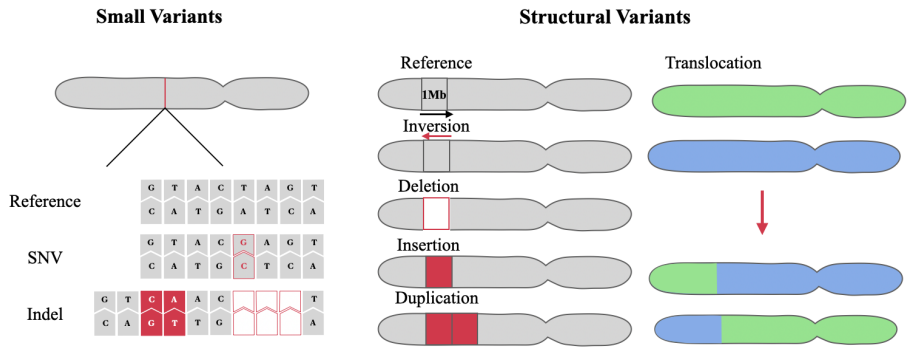


Figure 1.1: Small and structural variants. On the left are the small variants like single nucleotide variants (SNVs), insertion and deletion (indel). On the right, examples of up to 1Mb changes like inversion, deletion, insertion, duplication and translocation that constitute structural variants are shown. In each example, the top chromosome is the reference and the variation is highlighted and displayed below.

The abundance of small variants is much higher than the structural variants. About 85 million SNPs compared to 69000 structural variants found in the human genome (**Consortium et al. 2015**). These mutations can affect coding and non-coding regions or splice sites of the genome. They can be inherited or occur *de novo* in the germline. Many of these mutations are linked to various diseases, as they can alter protein structure or function, disrupt cellular processes, and contribute to disease phenotypes (**Visscher et al. 2012**).

Over the last decade, the genomics field has rapidly advanced. The development and application of large-scale next-generation sequencing (NGS) such as whole genome (WGS) and whole exome (WES) sequencing, has significantly expanded the capacity for comprehensive analysis of genetic variation. WGS is the method that sequences the entire genome of an organism including coding and non-coding regions, and captures all genetic variations. While WES is another approach focused solely on the exome sequencing or coding regions of the genome. These methods have both advantages and limitations, that should be taken into account. WES is more cost-effective than WGS and useful for identifying the mutations that affect the protein function. On the other side, it misses the variations in non-coding regions that can also impair gene regulation and cause the disease. Additionally, it's less effective in finding structural variants in comparison to WGS. To further expand our understanding of genetic variation, large-scale initiatives like the Exome Aggregation Consortium (ExAC) have emerged. ExAC contains the exomes of over 60,706 individuals, providing an extensive catalog of genetic variants across the whole exomes (**Lek et al. 2016**). With the integration of genome data, ExAC evolved into the Genome Aggregation Database or shortly gnomAD. gnomADAD is the largest public open-access human genome allele frequency reference database, which contains exome sequencing data from 730,947 individuals and genome sequencing data from 76,215 individuals.

Each individual's data is annotated with the population information. gnomAD includes data from various populations such as African, Latino, East Asian, South Asian, European, and others. The collected sequencing data is computationally processed to identify variants like SNPs, insertions, deletions, and other types of genetic variations. Up to now, it houses 786,500,648 single nucleotide variants, 122,583,462 InDels and over 1.2 million genome-level structural variants from more than 807162 individuals (**gnomAD 2024**). For each variant, the allele frequency is calculated, where the number of times the variant appears is divided by the total number of alleles observed at that position in the population. Researchers use this database to find threshold levels of variant

frequency within and across different populations. The knowledge about these levels help in understanding whether a variant is common or rare globally or only within specific populations (**Karczewski et al. 2020**).

While gnomAD provides valuable information on the frequency of genetic variants, it is not sufficient to determine which variants might be disease-causing. This is because population data alone lacks clinical and phenotypic information. For this purpose, patient data is essential. Sequenced patient data allows researchers to prioritize genetic variants with potential associations with specific diseases. To do prioritization, patient data is compared with population data (control) such as provided by gnomAD. This comparison involves the frequency analysis to examine whether a variant is more frequent in patients with the disease compared to healthy controls. Variants that are rare or absent from the general population, but found in affected individuals may be prioritized for further study. Next, statistical analysis is applied to assess whether the frequency of a variant is significantly higher in patients with the disease. This is only possible when there is a sufficient amount of patient data available. This helps identify variants that are statistically associated with the disease. The patient information is also used in studying the inheritance patterns within families to see if a variant co-segregates with the disease, which helps confirm its potential causative role. Finally, these variants found through the comparative analysis might undergo downstream functional studies.

Recognizing the need for comprehensive clinical data to better understand genetic variants, led to the development of patient databases. They represent the archives of the reported variants from patients submitted by clinical testing laboratories, research laboratories, locus-specific databases, expert panels, and other groups. The largest patient variant database currently known is ClinVar (**Landrum et al. 2016**). It is maintained at the National Center for Biotechnology Information (NCBI) within the National Library of Medicine and National Institutes of Health. Submissions to ClinVar must include a description of the variant(s), the interpreted condition, the clinical significance, an optional mode of inheritance, and supporting evidence.

Variants in ClinVar are classified based on the available evidence, including genetic studies, population frequency data, computational predictions, functional assays, and clinical observations. Pathogenic variants have statistical association with disease in large studies, evidence of segregation with disease in families, functional studies showing deleterious effects on gene or protein function, and consistent clinical observations in affected individuals. Benign variants are those that have been found frequently in healthy individuals. Variants that are between,

due to factors such as low frequency, insufficient population data, lack of functional studies or conflicting evidence are classified as variants of uncertain significance or VUS. Currently, ClinVar stores over 4 million submitted records and 2,966,675 genetic variants (**Landrum et al. 2016**; **ClinVar Miner 2024**). Of these, approximately 1,527,893 variants are VUS (**Figure 1.2**).

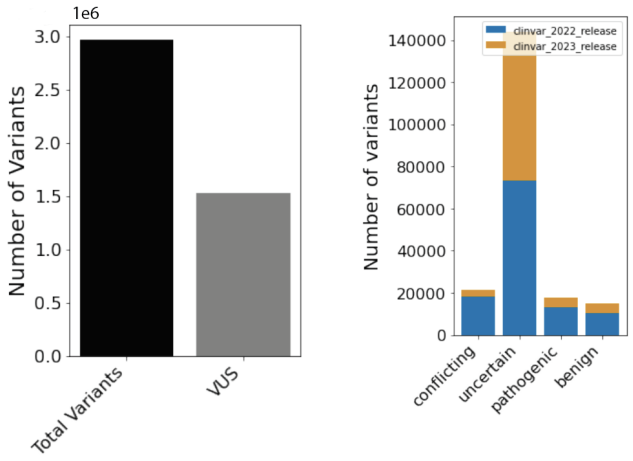


Figure 1.2: The variant distribution in ClinVar database. The bar plot on the left displays the total number of all variants (black) and the number of VUS (Variants of Uncertain Significance) (gray). On the right, the bar plot illustrates the distribution of various types of variants present in ClinVar for both the 2022 and 2023 versions.

Despite the remarkable advancements in sequencing technologies and increased data availability for research, the main challenge persists. The vast majority of variants remain poorly characterized. This number continues to grow exponentially each year, posing significant challenges for research and clinical practice. The accumulation of uncharacterized variants without corresponding progress in variant interpretation limits our understanding of the mechanism of diseases. Consequently, clinicians may struggle to interpret genetic test results, leading to potential delays in diagnosis and the development of precision medicine, where the treatment is tailored to each patient. For patients, this uncertainty can potentially result in anxiety, unnecessary treatments and missed opportunities for early intervention. Thus, why is it so challenging to characterize the variant effect?

To address this question, it is important to understand the underlying reasons behind the complexity of variant interpretation. Therefore, I will discuss them in the next subsection.

1.1.1 Factors contributing to the complexity of variant interpretation

The complexity of variant interpretation can be attributed to three primary reasons. First, the architecture of most diseases is highly complex, involving multiple genes. This means that a single variant may not have a straightforward impact on disease risk. Instead, its effect might be modulated by other factors like the interactions between gene products, the presence of other genetic variants or environmental factors. Even in Mendelian disorders, where one gene is the primary cause of the disease, the severity, onset and progression of a disease can still be influenced by additional genetic factors.

Second, many uncharacterized variants that occur infrequently in the human population, known as rare variants present a significant challenge. These variants are carried by only a small number of individuals. Additionally, every healthy individual on average carries about 60 *de novo* mutations (DNMs) that arise spontaneously and are not inherited from either parent, so-called ultra-rare variants can be extremely difficult to interpret (**Figure 1.3**). This low frequency of rare and ultra-rare variants results in small sample sizes, reducing the statistical power of tests. Statistical power refers to the ability of a test to detect a true effect when it exists. For instance, if a rare variant is present in only 0.1 % of the population in a study of 1000 participants, meaning that only 1 person will carry that variant. The statistical power will be too low because the small sample size reduces the distinction between true association from random fluctuations in data. With too few individuals carrying a rare variant, it becomes nearly impossible to apply standard statistical tests (e.g. Chi-square, Fisher's exact tests) effectively.

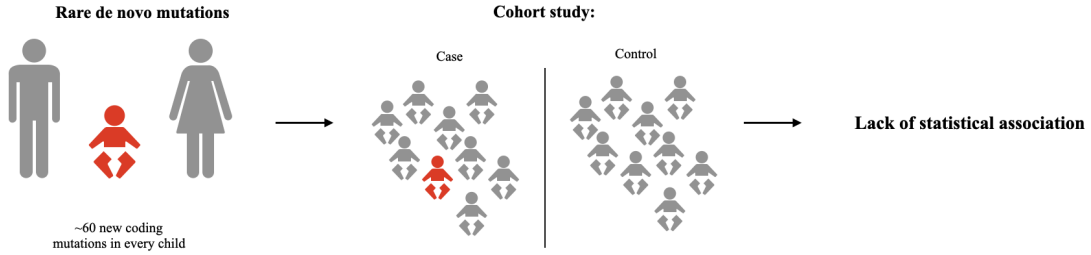


Figure 1.3: Schematic representation of the challenge in statistically associated *de novo* variants. Every healthy individual on average carries about 60 new coding mutations, most of them being ultra-rare in the human population. The low occurrence of these mutations makes it hard to do statistical association and discrimination of pathogenic from being variants.

Third, the impact of genetic variants on protein function can vary widely, spanning from benign alterations with no distinct effect to mu-

tations that cause severe dysfunction of the protein or disease (**Figure 1.4**). The traditional view on the mutation effect destabilizing the protein and leading to the loss-of-function (LoF) has evolved. For example, nonsense mutations possess a premature stop codon that leads to the truncated version of the protein, often misfolded or destabilized. The effect of the mutation might cause a severe phenotype. For example, a nonsense mutation in the DMD gene results in dysfunctional protein mutant Glu1157TER, causing muscular degeneration and Duchenne Muscular Dystrophy (DMD) (**Bulman et al. 1991**). Likewise, frameshift mutations, which result from deletions or insertions and alter the gene's frame can cause a total LoF due to extensive missense sequences followed by premature termination.

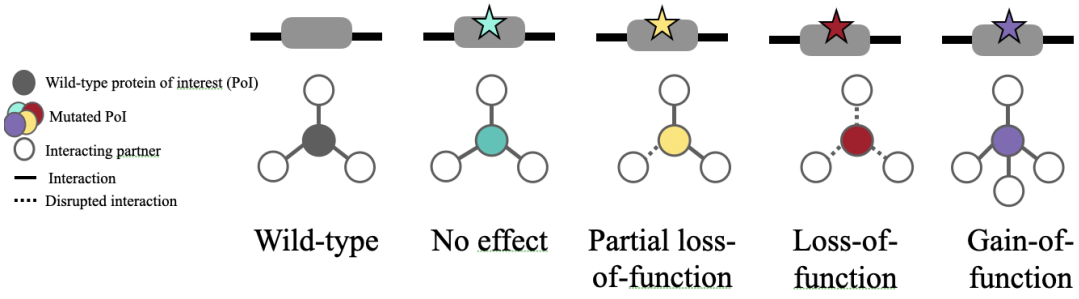


Figure 1.4: Overview of the various effects of mutation on PPIs. Mutations can have various effects: they may have no impact, destabilize and unfold the protein, cause a gain-of-function effect or partially affect protein function by disrupting PPI.

Although these types of mutations are known to be most detrimental and cause the disease, the reality is more complex, as in the case of the different mutations in the same gene causing distinct clinical outcomes (**Zhong et al. 2009**). This complexity arises because the gene and gene products do not function in isolation but in constant interactions with each other building interactome networks (**Vidal et al. 2011**). Goh et al (2007) assumed that some mutations possess a partial loss of function perturbing the interactions within this complex network (**Goh et al. 2007**). The studies suggest that about 30-60 % of all pathogenic missense variants are destabilizing, whereas up to 30 % of pathogenic missense mutations disrupt PPI, affecting some while leaving the others intact (**Sahni et al. 2013**). Given this information, how can we efficiently use it in variant characterization and learn about the potential mechanism of the disease?

1.1.2 Comparative PPI profiling as the strategy to interpret variant effect

The previous studies highlight the potential of using protein-protein interaction (PPI) data to interpret variant effects (**Zhang et al. 2009**; **Vidal et al. 2011**; **Sahni et al. 2013**). These interactions are represented as graphical networks, where proteins are displayed as "nodes" and the interactions between them as "edges". One promising experimental strategy that leverages this PPI network-based data is edgotyping, aimed to systematically characterize variants and reveal molecular mechanisms potentially underlying disease (**Sahni et al. 2015**; **Siwei Chen et al. 2018**; **Wierbowski et al. 2018**; **Starita et al. 2018**; **Fragoza et al. 2019**). The idea is based on testing and comparing the effect of benign, pathogenic, and uncharacterized mutations within a protein on the protein's interactions with its binding partners relative to the wild-type interactions. This comparison helps to identify the "edgetic" effects, where a mutation might disrupt some interactions while leaving the others intact. Thus, by comparing the obtained PPI profiles of benign, pathogenic, and uncharacterized variants we can predict the pathogenicity of this variant based on whether the interactions that are perturbed by the uncharacterized variant are similar to the interaction perturbations observed for the pathogenic variant. Moreover, the obtained perturbation data might be informative about potential mechanistic causes of the disease, making the strategy powerful at elucidating functional consequences of variants on PPIs and insight into variant contributions to the disease that go beyond traditional sequence-based studies.

Available human PPI datasets

To perform edgotyping, access to comprehensive and reliable protein interaction datasets is crucial. Over the past 15 years, high-quality human PPIs have been generated using large-scale approaches such as yeast two-hybrid (Y2H) and affinity purification coupled with mass spectrometry (AP-MS). They have become particularly instrumental in mapping PPIs and generating large-scale reference protein interactome datasets (**Rolland et al. 2014**; **Luck et al. 2020**; **Huttlin et al. 2021**). Rolland et al (2014) made a significant input into this field by presenting a broadened version of the human interactome, HI-II-14, consisting of about 14000 distinct protein interaction pairs determined and confirmed by three binary PPI assays (**Rolland et al. 2014**). This available dataset has been instrumental in the research focused on variant characteriza-

tion. They further investigated the overall biological relevance of this PPI dataset assessing mutations associated with human disorders compared to common variants that showed no functional consequences on biophysical interactions. They showed that more than 55 % of the 107 tested PPIs were perturbed by at least one disease-associated variant. For example, the A129T mutation in the AANAT protein is known to be associated with delayed sleeping phase syndrome. It specifically disrupted the interaction with BHLHE40 involved in the regulation of circadian rhythm. Another study utilized HI-II-14 and overlapped it with nearly 2000 mutations and identified 298 disruptive variants affecting almost 700 human protein interactions (**Fragoza et al. 2019**).

In 2020, the human reference interactome or HuRI unveiled the largest binary interaction map of human proteins using the Y2H approach. The HuRI project employed the Y2H technique, where two proteins are co-expressed in yeast cells if they physically interact, resulting in a total of 3 billion individual tests. This monumental effort generated a dataset of over 50000 high-confidence binary interactions between approximately 17000 human proteins. The depth and the scale of this study significantly enhanced our understanding of the human interactome and provided valuable datasets for elucidating the functional impact of patient variants on protein-protein interactions (**Luck et al. 2020**). Luck et al (2020) also showcased the application of HuRI in elucidating the mechanistic effect of missense variants on PPIs within specific disease contexts. Mutations in PNKP have been associated with microcephaly, seizures and developmental delay. They showed that the pathogenic mutation Glu326Lys in PNKP disrupted the interaction with TRIM37 predominantly expressed in the brain. Here, these studies demonstrated that systematically generated human interactome maps may significantly help in variant characterization.

In parallel, the Bioplex project generated a reference human interactome using the AP-MS technique. This study involved systematic protein purification and bound potential binding partners from cells, followed by a mass spectrometric analysis of protein complexes. BioPlex mapped about 120000 direct and indirect protein interactions (**Huttlin et al. 2021**; **Huttlin et al. 2017**). In addition, it was also employed for variant characterization, identifying how mutations affect not only direct interactions but also complexes relevant to diseases. As a result, it offers a broader view of variant functional impact on protein complexes compared to Y2H. However, Y2H-detected interactome might be more useful for studying variant effects, as it provides binary protein interactions essential for comparative PPI profiling. This approach tests how mutated protein affects each specific interaction, enabling the creation

of mutation profiles and their comparison with those of the wild-type protein and its partners (**Idrees et al. 2024**).

Application of edgotyping strategy

Several successful attempts were made to perform this approach (**Sahni et al. 2015**; **Siwei Chen et al. 2018**; **Wierbowski et al. 2018**; **Starita et al. 2018**; **Fragoza et al. 2019**). For example, Sahni et al (2015) generated interaction profiles for 460 mutant proteins and their 220 wild-type counterparts and found 521 perturbed interactions out of 1,316 PPIs using the yeast two-hybrid (Y2H) interaction assay. This huge experimental effort led to the identification of 197 mutations, where 26% identified as complete loss of interaction, 31% as edgetic and 43% had no change in PPIs. Later Fragoza et al (2019) employed the same assay and identified 298 out of tested 1676 missense population variants that disrupted 669 human PPIs. They also used follow-up experiments to further elucidate the effect of mutation on protein function. Taken together these attempts showcase how shared disruption profiles can be used to prioritize candidate disease-associated mutations.

Current challenges in edgotyping

While this approach holds significant potential for addressing the issue of uncharacterized variants, it is still too expensive and laborious, if it is entirely based on experiments given the amount of VUS that needs to be characterized. While current tools like PolyPhen-2 and MutPred2 predict variant pathogenicity primarily use metrics such as conservation score or sequence-based features related to protein structure and function fail to capture the effect of mutations occurring in less conserved but yet functional regions or rare and ultra-rare variants with low conservation scores (**Sunyaev et al. 1999**; **Adzhubei et al. 2010**; **Livesey et al. 2022**). The recently developed AlphaMissence tool excels in performance but also shows less effectiveness for variants in these regions (**Cheng et al. 2023**). Given these limitations, comparing edgetic profiles of benign, pathogenic with VUS variants might not always be sufficient to identify functional variants potentially contributing to the disease. How can PPI profiling be improved for more effective variant characterization?

To predict the variant effect on PPIs, one needs ideally to know the exact residues that constitute the protein interaction interfaces (**see sections 1.2 and 1.3**). Access to this information would be extremely useful, as it helps pinpoint exactly where and how a mutation might

disrupt an interaction and elucidate the mechanistic effect and potential impact of a variant on disease development. This assumption is supported by the study, where they reported a significant enrichment of disease mutations found on the PPI interfaces (**wang_2012**). Although we have protein interaction datasets available, they carry only binary information, while the information on PPI interfaces is currently missing. Various experimental approaches such as X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, cryo-electron microscopy (cryo-EM) and protein fragmentation exist to detect PPI interfaces at different resolutions (**Martino et al. 2021**). However, experimental methods are labor-intensive and time-consuming. Indeed, only a small fraction of interactions, about 4 % in the HuRI dataset have solved structures (**Luck et al. 2020**). Given the limitations of experimental studies, computational methods to predict PPI interfaces have gained traction in recent years. The idea is to increase the predictive power and map PPI data with putative PPI interfaces that further accelerate the experimental validation of the putative PPI interfaces (**see section 1.4**). Finally, this information will be used for PPI profiling described earlier.

1.2 Modular architecture of proteins

The prediction of PPI interfaces requires an understanding of protein architecture to identify functional sites along with databases of known functional sites to enhance the accuracy of these predictions. This will be discussed in this section.

Proteins are complex molecules that play a crucial role in cellular biological processes. Since the advent of molecular biology, we learnt that proteins do not function in isolation, but in constant interactions with one another or other molecules (i.e. RNA, DNA) forming complex networks. These interactions are mediated by PPI interfaces formed by specific regions within protein sequences, widely known as functional modules (**Campbell et al. 1991**). These modules broadly can be categorized into defined and undefined structures. The defined structures, commonly known as the globular domains, are the regions in protein sequences that often independently fold into a stable tertiary protein structure (**Copley et al. 2002**; **Björklund et al. 2005**). Those regions that lack a defined structure are termed intrinsically disordered regions or IDRs, where short linear motifs (SLiMs) are typically found (**Tompa et al. 2014**; **Davey et al. 2011**; **Davey et al. 2012**). These two types of functional modules will be explained further in the following

sections.

The modularity of the proteins is a crucial aspect of protein evolution and functionality (**C. Vogel et al. Year**; **Han et al. 2004**). This modularity allows combining different modules to make proteins with multiple properties and functions, facilitating the diversity of new traits and adaptation to environmental changes (**Apic et al. 2001**). Around 65-70 % of proteins in eukaryotic organisms are composed of multiple modules in their proteomes (**Han et al. 2004**). One prominent example is the well-characterized Nuclear factor NF-kappa-B p105 subunit or NFKB1, a multifunctional hub protein and transcription factor involved in various cellular processes, such as transcriptional regulation, immune response, cell proliferation and survival (**Gilmore 2006**; **Hayden et al. 2008**). It has been implicated in a broad range of cancers, neurodegenerative diseases, and inflammatory and autoimmune diseases (**Gilmore 2006**; **Hayden et al. 2008**; **Taniguchi et al. 2018**). The N-terminus of 968 amino acid-long protein starts with the Rel homology domain (RHD), followed by Ankyrin repeats and Death domain (DD) at the C terminus (**Williams et al. 2001**; **Glover 2004**; **J. Wang et al. 2023**). The disordered parts of the protein harbor many known motifs such as nuclear export and nuclear localization signals, docking and kinase modification motifs (**Koonin 1996**; **Chen et al. 1996**; **Rodríguez et al. 2000**; **Hsu 2007**). Thus, the modularity in NFKB1 enables it to interact with many different partners, function in various cellular processes and exemplify the complexity of the phenotypes that can arise from the interplay of functional modules (**Figure 1.5**).

1.2.1 Folded domains

Biological role of domains

The foundational understanding of protein domains began with the work of structural biologists Linus Pauling and Robert Corey in the 1950s. Their research identified alpha helices and beta sheets as secondary structures within proteins. Wetlaufer and Ristow (1973) introduced the concept of protein domains or functional modules in their review of X-ray crystallography studies of enzymes like lysozyme and immunoglobulins (**Blake et al. 1965**; **Freedman et al. 1966**). They associated domains with the regions, typically ranging from 50 to 350 amino acids in length that are capable of folding autonomously. This understanding was facilitated by the development of experimental methods like crystallography and NMR, which accelerated the identification and classification of these protein modules, including commonly found domains in proteome such

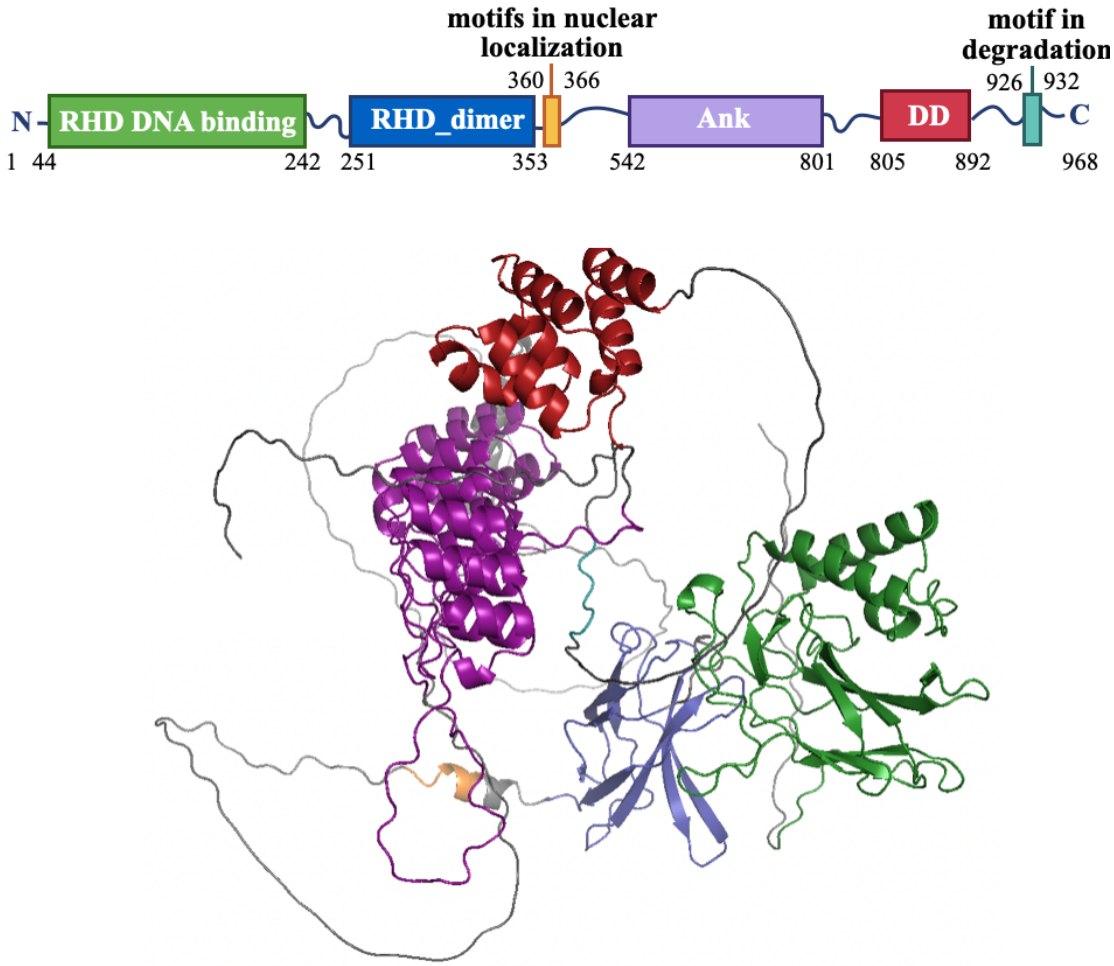


Figure 1.5: Modularity of Nuclear factor NF-kappa-B p105 subunit. Domain and motifs as functional modules schematically illustrated in NFKB1. The top panel shows the modularity of NFKB1. The numbers above and below the boxes denote the boundaries of domains. The bottom panel displays the full-length structure of NFKB1 predicted by AlphaFold. Domains and motifs in the structure are colored according to their colors in the top panel. RHD stands for Rel Homology Domain, Ank stands for Ankyrin repeats, and DD stands for death domain.

as WD40, SH2, SH3, ANK, RING, PH and PDZ (**Copley et al. 2002**).

Domains fold into three-dimensional (3D) structures to achieve thermodynamic stability, positioning the hydrophobic residues in the protein core while exposing hydrophilic residues on the surface (**Dill et al. 2012**). This ensures that the native conformation is the most energetically stable for the domain. Domains form the functional units of a protein, enabling it to interact with partners and perform cellular functions. For instance, the protein KLHL41 is involved in the ubiquitin-proteasome system, which regulates protein turnover and degradation, maintaining various biological processes like muscle development and the function (**Ramirez-Martinez et al. 2017**; **Yuen et al. 2020**). KLHL41 consists of three main domains: the Broad-complex, Tramtrack, Bric-a-brac (BTB) domain, BTB and C-terminal Kelch (BACK),

and Kelch repeats (**Figure 1.6**). The Kelch repeats of KLHL41 form a beta-propeller structure that recognizes substrates such as nebulin (NEB), a giant muscle protein that acts as a molecular ruler for filament length and regulates actin-myosin cross-bridge cycling during skeletal muscle contraction (**Yuen et al. 2020**). Upon binding, KLHL41 forms a complex with NEB, while the BTB domain of KLHL41 directly binds with cullin 3 (Cul 3), a scaffold protein in the Cullin-RING ubiquitin ligase (CRL) complex, and can dimerize with itself to provide more stability to the complex. Additionally, the BACK domain at the C-terminus of KLHL41 supports and stabilizes the complex (**Stogios et al. 2004**; **Dhanoa et al. 2013**; **Gupta et al. 2014**). Once the substrate is formed, the ubiquitin molecules are transferred to the substrate subunit, marking the substrate protein for the degradation by the proteasomal system. This case illustrates how linking different domains together in one polypeptide chain allows KLHL41 to maintain protein homeostasis.

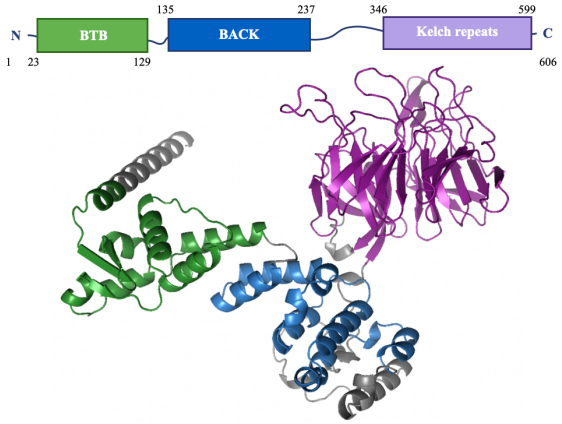


Figure 1.6: Domain architecture of Kelch-like protein 41 (KLHL41). The top panel shows the schematic domain organization of KLHL41 (not drawn to scale). The numbers above and below the boxes denote the boundaries of domains. The bottom panel shows the putative full-length structure of KLHL41 as predicted by AlphaFold. Domains and motifs in the structure are colored according to their colors in the top panel. BTB stands for the Broad-complex, Tramtrack, Bric-a-brac domain and BACK - for BTB and C-terminal Kelch domain.

While some domains can achieve stability independently or through dimerization, others require assistance from zinc and metal ions or disulfide bridges. For instance, the zinc finger domain maintains its conformation by binding to zinc ions. These zinc ions typically interact with cysteine and histidine residues, acting as an anchor that reduces the protein chain flexibility and supports the stable 3D structure (**Berg et al. 1997**; **Klug 2010**). This stabilization is important for protein functions such as DNA binding and gene expression. A good example is the IKZF1 protein, also known as Ikarios, a zinc finger protein and tran-

scription regulator that plays a crucial role in lymphocyte differentiation and function. It contains four C2H2-type zinc finger domains at the N terminus that bind to zinc ions. The stabilized structure of a protein interacts with DNA sequences in the promoter regions of targeted genes and regulates their transcription. Different combinations of protein domains exist widely across proteomes due to natural selection, acting on these modular units to create diverse molecular machinery (**Doolittle 1995**).

Gene duplication and shuffling by recombination are likely to be the driving forces of protein evolution and the complexity of the proteome. While gene duplication leads to the emergence of similar domains occurring in unrelated proteins, recombination enhances versatility and allows proteins to specialize in specific cellular functions tailored to an organism's needs (**Bagowski et al. 2010**). For example, the PDZ domain is a 90-100 residues long structurally conserved module, found in a vast array of proteins involved in diverse signaling pathways and cellular polarity (**Harris et al. 2001**; **Lee 2010**). About 270 PDZ domains are distributed over 150 proteins (**Wang et al. 2010**; **Velthuis et al. 2011**). Despite their conserved structural fold, these domains exhibit sequence divergence that contributes to functional specificity. Thus, PDZ domains in the protein PSD-95 recognize C-terminal motifs on its target proteins, whereas the PDZ domain in cell polarity protein PAR6 was shown to interact with internal ligands and other PDZ domains can form homodimers (**Kornau et al. 1995**; **Zhang et al. 2009**; **Fouassier et al. 2000**).

Protein domain databases

Previously, a big contribution to the discovery of protein domains was done by sequencing projects. This effort helped to identify the conserved regions across different proteins. With the power of bioinformatics, the domains became identifiable using Hidden Markov models (HMMs). HMM is a statistical model used to classify protein families based on multiple sequence alignments (MSA) and detect sequence homology for the identification of conserved regions within proteins (**Bystroff et al. 2008**). Thus, HMM became the main approach to collecting the data and generating domain databases.

For example, the Protein families database (Pfam) and Simple Modular Architecture Research Tool (SMART) computed HMMs to build protein and domain families based on the sequence similarity (**R. D. Finn et al. 2014**; **Schultz et al. 1998**; **Letunic et al. 2021**). While the SMART database has manually curated sequence alignment

which helped to define the domain boundaries more precisely, Pfam employs the automated approach and covers a broader range of domains (**Paysan-Lafosse et al. 2023**).

1.2.2 Intrinsically disordered regions

Biological roles

In the mid-20th century, protein research primarily focused on folded and ordered proteins. Studies on enzymes, where the denatured proteins loose their catalytic activity, demonstrated the relationship between protein structure and function (**Northrop 1930**). It was assumed that a protein requires a native folded structure to perform biological functions. Thus, the protein structure-function paradigm was established, while the abundance and functional role of disordered regions in proteins in eukaryotes was unrecognized. However, unexpected behavior of proteins such as missing electron density in X-ray crystallography studies, increased sensitivity in the *in vitro* proteolysis experiments and solubility issues during protein purification processes led to the reassessment of the structure-function paradigm. Pioneering work by Dunker (2002) and Urevsky (2005) revealed that disordered regions are common in eukaryotic proteins. Further studies challenged the long-standing belief that protein functionality was strictly dependent on a well-defined and folded protein structure (**Tompa 2002**; **Dunker et al. 2002**; **Wright 1999**; **Iakoucheva et al. 2002**; **Uversky et al. 2005**; **Uversky 2014**). It was shown that disordered regions of many regulatory and signaling proteins can undergo disorder-to-order transitions upon binding to their targets, which adds a layer of regulatory control and allows for complex interactions (**Uversky 2014**). As a result of these findings, the scientific community began to recognize the importance of protein disorder, leading to a significant shift in understanding protein biology. Consequently, the paradigm was shifted to the "disorder-function paradigm".

Intrinsically disordered regions (IDRs) lack persistent 3D structure under physiological conditions, continuously adopting the wide range of dynamic conformations and forming transient secondary structures (**Wright 1999**; **Tompa 2011**; **Davey et al. 2019**). These regions are abundant in eukaryotic proteins, with predictions indicating that they cover 30-40% of residues in their proteome (**Tompa 2012**; **Van Roey et al. 2012**). IDRs also significantly contribute to the diversity and versatility observed in organism evolution (**Davey et al. 2015**; **Babu et al. 2012**; **Weatheritt et al. 2012**). In addition, they are often found to overlap with post-translational modifications (PTMs), contributing

to functional versatility (**Tompa 2012**; **Tompa et al. 2014**). These modifications can alter the conformation, stability and interactions mediated by IDRs (**Uversky 2014**). Due to the dynamic behavior, IDRs are commonly involved in transient interactions regulating signal transduction processes (**Dyson 2005**; **Davey et al. 2019**). A crucial finding was that IDRs are enriched with functional interaction modules, such as short linear motifs (SLiMs) mediating different multivalent interactions, which will be discussed in the next section.

As IDRs play a significant role in signaling and cell regulation, they are tightly controlled, and mutations in disordered sites have been associated with human diseases, including cancer, diabetes, cardiovascular and neurodegenerative disorders (**Iakoucheva et al. 2002**; **Babu et al. 2011**). Vacic et al. (2012) investigated disease-causing missense mutations on ordered and disordered regions and compared them to neutral variants observed in healthy individuals without causing disease phenotypes. They found that over 20 % of pathogenic variants reside in intrinsically disordered regions and interfere with their functions. In addition, the study by Peng et al (2012) emphasizes the importance of understanding the context-dependent behavior of IDRs. They highlighted that the functional outcome of missense variants in these regions could vary depending on the cellular environment and interaction partners (**Peng et al. 2012**).

Despite the biological relevance of IDRs, only a small fraction of IDRs have been characterized (**M. Gouw et al. 2017**; **Davey et al. 2019**). Experimentally, defining disordered regions remains challenging. Due to the dynamic structures of IDRs, the use of sophisticated methods such as NMR, small-angle X-ray scattering (SAXS), circular dichroism (CD) or Förster resonance energy transfer (FRET) is required (**Felli et al. 2015**; **Holmstrom et al. 2016**). Moreover, these regions function in a context-dependent manner based on the cellular milieu including pH, PTMs, and the presence of other proteins (**Oldfield et al. 2014**; **Wright 2015**). These challenges necessitate integrative approaches that combine experimental data with computational predictions. As a result, various computational approaches have been developed to predict IDRs in proteins, leading to the generation of several databases containing putative IDRs and experimentally verified.

Databases of disordered proteins and tool to predict the disorderness

The DisProt is a comprehensive repository of experimentally verified entries of proteins or regions within proteins that lack a stable three-

dimensional structure under physiological conditions, where each entry is manually curated. The DisProt database annotates the disorder and molecular functions curated from experimental studies. More than 2,000 eukaryotic intrinsically disordered proteins (IDPs) and 6,000 IDRs are documented in this database.

IDRs possess distinctive characteristics that set them apart from structured regions. One notable characteristic is the enrichment in polar hydrophilic residues coupled with the depletion of hydrophobic amino acids that help to stay soluble and flexible in the disordered state, and incapable of forming sufficient interresidue interaction within a protein. To discriminate between ordered and disordered sequences, the Intrinsic Unstructured Protein Predictor tool (IUPred) developed the approach, where they calculated the likelihood of interaction formations using a statistical interaction potential (**Z. Dosztányi 2018**). These potentials are further used to assess each residue in the protein sequence to estimate their energies. Based on the energy, the residues estimated to have the most favorable energies are predicted to be ordered, while those with unfavorable energies are predicted to be disordered (**Mészáros et al. 2009**).

Recently, a new powerful tool AlphaFold2 (AF2) has emerged, predicting protein structures with accuracy comparable with experimental structures (**Jumper et al. 2021**). AF2 predicts a full-length protein structure generating the confidence score termed as Local Distance Difference Test (pLDDT). pLDDT score is calculated for each residue in the protein structure, where it ranges from 0 to 100. A high score means greater confidence in the accuracy of the prediction of a residue's position. Interestingly, the pLDDT was found to correlate with the disordering tendency, which can be used as a potential feature to predict disorder (**Wilson et al. 2022**).Another feature is the solvent-accessible surface area (SASA) of each residue is also correlated with the disorder propensity of residues. One study used both pLDDT and SASA and smoothed over a 20-residue window and outperformed IUPred2A, the latest version of the predictor tool in their study (**Akdel et al. 2022**). As AF has been used for other applications, they will be described in section 1.4.

1.2.3 Short linear motifs

Biological roles

Short linear motifs (SLiMs) represent dynamic functional sequences, ranging from 3-23 amino acids long. On average four residues are con-

served in the motif consensus sequence, but the remaining positions are completely variable (**Davey et al. 2012**). Motifs typically lie in IDRs or more rarely in disordered loops of structured regions and possess regulatory functionality such as directing ligand binding, providing docking sites for enzymes and targeting proteins to specific subcellular locations (**Davey et al. 2012**; **Van Roey et al. 2014**).

The concept of SLiMs appeared in the late 20th century. In 1980 Aaron Ciechanover, Avram Hershko and Irwin Rose identified degradation motifs or degrons that direct the target proteins to the ubiquitin-proteasome system for degradation. Their groundbreaking work earned them a Nobel Prize in 2004 and laid the foundation for discovering new motifs. In 1990, Tim Hunt identified targeting signals such as KDEL endoplasmic reticulum retention motif, and the positively charged nuclear and targeting sequences, while Pawson et al. (1986) discovered that Src domains recognize motifs within protein partners, the interactions with which regulate signaling pathways. These studies highlighted the importance of motifs in protein function and regulation, opening avenues for further exploration and discovery in molecular biology and cellular physiology.

The discovery and validation of SLiMs have been performed by various experimental methods such as traditional low-scale X-ray crystallography and NMR as well as high-throughput systematic approaches such as peptide microarrays, and phage display. Along with experimental discoveries, the computational approaches have also significantly advanced motif research. The motif detection techniques will be discussed in more detail in sections 1.4 and 1.5.

It is estimated that more than 100,000 binding motifs exist in the human proteome, with many being uncharacterized (**Tompa et al. 2014**). The discovered motifs are categorized into six classes based on their biological roles: ligand-binding sites, modification, targeting signals, degrons, docking and cleavage. Modification motifs include PTM sites like phosphorylation. Targeting signals like nuclear localization signals (NLS) are involved in protein trafficking to specific cellular compartments. Ligand-binding motifs interact with binding partners to form transient signaling complexes. Docking motifs facilitate substrate recognition by enzymes without affecting the active site of these enzymes. The cleavage motifs are recognized by proteases that cleave the protein at the cleavage site (**Van Roey et al. 2014**). Another functional type of motif is degron. Degrons, such as those, found in the protein AFF4 (**Figure 1.7**), are important for protein regulation. Specific ubiquitin ligases like SIAH1 recognize these motifs which tag the target proteins with ubiquitin molecules. This tagging process, known as ubiquitination

marks the protein for degradation by the proteasome system (**Oliver et al. 2004**).

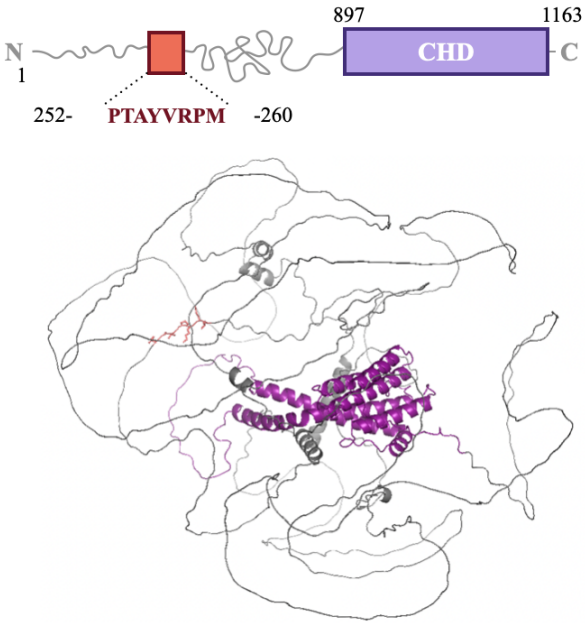


Figure 1.7: Degron motif on AFF4.
The top panel shows the schematic domain organization of AFF4 (not drawn to scale). The numbers above and below the boxes denote the boundaries of domains. AFF4 contains a degron motif that is recognized by ubiquitin ligase. The bottom panel shows the putative full-length structure of AFF4 as predicted by AlphaFold 2. Domains and motifs in the structure are colored according to their colors in the top panel. CHD stands for C-terminal homology domain.

Moreover, they mediate transient regulatory and signaling interactions involved in biological processes like cell signaling, protein homeostasis and cell cycle. For instance, the 14-3-3 binding motif facilitates the interaction of diverse proteins with 14-3-3 domains, thereby regulating their subcellular localization and activity of 14-3-3 proteins. Another example is the SH3-binding motif (PXXP), found in numerous signaling proteins, which mediates interactions with SH3 domains of other proteins, facilitating the assembly of signaling complexes in response to extracellular stimuli (**Davey et al. 2012**; **Van Roey et al. 2014**).

Additionally, SLiM mimicry can be used by viruses to interfere with the host cellular machinery and thereby repurposing the host cell for pathogen reproduction (**Davey et al. 2011**; **Uyar et al. 2014**). For example, the Nsp3 protein of Eastern equine encephalitis virus (EEEV), contains the motif LITFD that mimics the classical clathrin box motif. This mimicry allows Nsp3 to interact with the beta-propeller repeat of the N-terminal domain of clathrin (CLTC). This interaction disrupts clathrin-mediated receptor trafficking and interferes with the signaling processes, potentially suppressing antiviral signaling or altering cellular

functions to create a more favorable environment for viral replication (**Mihalič et al. 2023**).

As opposed to globular domains, SLiMs are short functional peptides and take up a very small sequence space. Consequently, IDRs can be densely packed with multiple SLiMs, which can sometimes overlap and act as regulatory switches. There are different switch mechanisms. One of the mechanisms is switching the specificity of protein to its binding partners like modification-dependent modulation of the intrinsic affinity of the motif. The protein integrin beta 3 is the cell surface receptor involved in cell adhesion and cell signaling (**Tadokoro et al. 2003**). The NPxY motif in the disordered tail of the integrin beta 3 subunit preferentially interacts with the PTB domain and membrane proximal region of talin necessary for the integrin activation (**Wegener et al. 2007**). However, phosphorylation of the motif, particularly at position Tyr747 switches the specificity to PTB of Dok1. Dok1 prefers to bind exclusively to the central motif and does not interact with the membrane-proximal region of the integrin tail necessary for activation. Therefore, this mechanism ensures the control over integrin-mediated cellular processes (**Oxley et al. 2008**).

Computational and experimental studies have shown that pathogenic mutations in disordered regions often affect SLiMs. Uyar and colleagues (2014) performed a proteome-wide analysis of disease-associated mutations with a focus on SLiMs. Here, they utilized the mutation data from healthy and patient individuals reported in databases such as Catalog of Somatic Mutations In Cancer (COSMIC), 1000 Genomes Project, and Online Mendelian Inheritance in Man (OMIM), respectively (**Consortium et al. 2015**; **Forbes et al. 2011**; **Hamosh et al. 2005**). Next, they mapped these mutations on SLiM derived from the experiment and putative SLiMs using the IUPred tool and compared the distribution of pathogenic and neutral mutations. The analysis revealed that disease-related mutations are significantly enriched on SLiMs within intrinsically disordered regions (**Uyar et al. 2014**). Additionally, mutations within SLiMs can disrupt motifs or create new ones. The study experimentally showed that pathogenic mutants formed dileucine motifs that often lead to clathrin-binding that underlies disease aetiology (**Meyer et al. 2018**).

This accumulated evidence highlights the importance of SLiMs as a key aid to understanding the molecular mechanisms in diseases and underscores the need to integrate SLiM analysis into variant characterization studies.

Motif databases

While Pfam and SMART are valuable for predicting domain-involving interfaces, motif databases can help identify potential SLiM-mediated interfaces. The Eukaryotic Linear Motif (ELM) is a comprehensive database developed by Toby Gibson and colleagues in the early 2000s. The ELM database provides researchers with a catalog of manually curated and experimentally annotated validated SLiMs and tools for motif prediction with the main focus on annotation and detection of SLiMs (**Puntervoll et al. 2003**). Each record provides extensive information on the motif sequence pattern, functional role, interaction partners, biological processes it influences and experimental evidence. Additionally, the database has a search interface that allows users to query the motif based on the sequence pattern, protein identifier, and species.

The ELM database categorizes motifs into functional types, classes and instances. There are 6 functional types of SLiMs: ligand-binding (LIG, e.g. WW1 binding motif), modification (MOD,e.g. CK1 phosphorylation site), targeting (TRG, e.g. NLS classical nuclear localization signal), docking (DOC, e.g. USP7-binding motif), degradation or (DEG, e.g. Siah binding motif) and cleavage or (CLV, e.g. NRD cleavage site). These types are grouped into 356 ELM classes based on the binding domain of a partner, specific sequence characteristics, targeted subcellular localization and other functional properties (**Kumar et al. 2024**). These classes incorporate 4283 individual ELM instances manically curated from 4274 scientific publications and 2749 motif-partner interactions (**Kumar et al. 2024**). Each instance has annotated details on the evidence like the experimental method used to determine and characterize the discovered motif (**M. Gouw et al. 2017**). ELM curators systematically described each ELM class using a regular expression (RegEx) format to define the key residues important for the binding affinity and specificity of the motif (**Davey et al. 2011**). These regular expressions also capture the conservation pattern of different motif types and, therefore, can be used in the prediction of putative motifs.

1.3 Domain-motif interfaces

Current understanding of protein-protein interaction interfaces

Protein interaction interfaces are formed through the interaction of protein modules, mainly globular domains and motifs. For example, the binding between two globular domains is termed domain-domain inter-

face (DDI). DDIs involve multiple contacts and are characterized by a high binding affinity, which contributes to the stability of protein interactions (**Nooren 2003**). DDI interactions aid in stabilizing the formation of protein complexes and are often involved in enzymatic activity, cell signaling, cell adhesion and other cellular events.

Later, researchers found that in addition to DDIs, protein domains can recognize SLiMs forming a domain-motif interface or DMI (**Dyson 2005**; **Babu et al. 2012**; **Tompa 2012**; **Davey et al. 2012**). DMI-mediated interactions are weaker and more transient, playing a role in major biological processes such as signal transduction, protein targeting to cellular compartments and protein homeostasis (**Schreiber et al. 2009**; **Zhou 2012**). Therefore, maintaining these DMI interactions is crucial, as their disruption can potentially lead to the disease (**Arimura et al. 2000**; **Uyar et al. 2014**). Despite the importance of DMIs, they are significantly underrepresented. Due to the transient nature of these interactions, it is hard to detect using traditional experimental approaches described in section 1.5. Tompa et al (2014) estimated the number of motifs in the hundreds of thousands or even millions. Therefore, the last two decades have seen a tremendous rise of interest in SLiMs interface-mediated PPIs in different research fields like structural biology, systems biology and bioinformatics.

In my thesis, I will focus only on the systematic prediction of DMIs followed by experimental validation and will use this information in comparative PPI profiling as the strategy for efficient variant characterization.

Functional significance of Domain-Motif interfaces in cellular processes and disease

In this section, I will describe several examples highlighting the functional role of DMI interactions in biological processes and their implications for the disease.

A notable example of these is the degron motif with the pattern PxAxVxP, where x represents any amino acid) is found in the target protein AFF4. This protein plays a critical role in transcription regulation and chromatin remodeling and it is a core component of the super elongation complex (SEC). SEC facilitates the efficient synthesis of mRNA transcripts by RNA polymerase II (RNAPII) during transcription elongation (**Lin et al. 2010**; **C. Luo L. et al. 2012**). This protein also helps to recruit RNAPII to gene promoters and overcome the transcriptional pausing. This activity is crucial for ensuring proper gene expression profiles and supporting cellular function (**Lin et al. 2010**). The degron

motif of AFF4 is recognized by the substrate-binding domain (SBD) of E3 ubiquitin ligase, SIAH1.

SIAH1 is the central component of a multiprotein Er ubiquitin ligase complex and essential for protein level regulation within the cell. It has been implicated in the regulation of programmed cell death. In some studies, it has been identified as a tumor suppressor as it can degrade the oncogenic proteins. This helps to prevent tumor formation and progression. The recognition of AFF4 by SIAH1 has been previously functionally annotated (**Oliver et al. 2004**). Upon binding this motif forms a beta strand parallel to the beta-sandwich fold of the substrate binding domain (SBD) of SIAH1. This interaction is known as the beta augmentation mechanism. When the SBD of SIAH1 contacts the degron of AFF4, it facilitates the ubiquitination of AFF4. Then this tagged protein is degraded by the proteasome complex (**Figure 1.8**). This biological process is important for maintaining homeostasis in the cell by removing damaged and misfolded proteins and regulating protein levels within the cell (**Santelli et al. 2005**).

While the mechanism of the interface between these proteins has been annotated, the exact mechanism underlying the development of these disorders is poorly understood, and many mutations found on these interfaces remain uncharacterized. For example, the Met260Thr variant, where methionine is mutated to threonine within the motif of the previously mentioned AFF4. The mutation was found in the patient with a rare NDD called CHOPS syndrome and reported in Clinvar as VUS. However, the diagnosis of CHOPS syndrome, caused by this rare mutation is complicated. The limited number of documented cases makes establishing diagnostic criteria and developing personalized treatment difficult. Using our approach we know that the mutation is sitting on the motif of AFF4 and might perturb the interaction with SIAH1. The disruption of interaction may lead to the stabilization and accumulation of AFF4 and cause developmental abnormalities characterizing the disease.

Another example of the domain-motif mediated interaction is the interaction between the 14-3-3 domain proteins and phosphorylated ligand motifs **Figure 1.9** on the target proteins (**Grozinger et al. 2000**; **M. J. Wang K. et al. 2000**). YWHAG (14-3-3 protein gamma) is one of the proteins possessing a 14-3-3 domain which recognizes phosphorylated serine residues within the RAQSSP, RTQSAP and RKTASEP consensus motifs of histone deacetylase 4 (HDAC4). This interaction is known and the motif binding to 14-3-3 proteins was first described in 1997 by Yaffe et al. YWHAG is an adaptor protein localized in the cytoplasm. It belongs to the 14-3-3 protein family involved in signal trans-

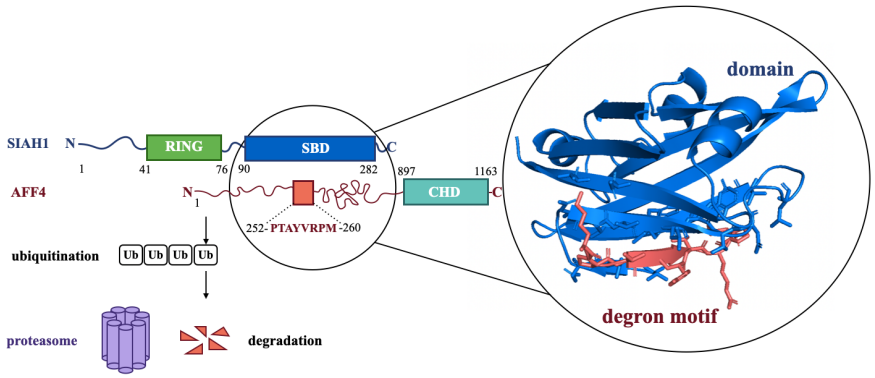


Figure 1.8: The mechanism of interaction between SIAH1 and its target partner AFF4.
Substrate-binding domain (SBD) binds to the degron motif on AFF4, which leads to the ubiquitination of AFF4. Tagged protein is further degraded by the proteasomal system (Oliver et al., 2004). CHD stands for C-terminal homology domain. The structure of the interface is shown as predicted by AF2.

duction, protein localization, cell apoptosis and cell cycle. This protein plays a crucial role in signaling pathways by binding to the phosphorylated motifs of its interacting partners. One of these proteins is HDAC4, a transcriptional regulator, which deacetylates lysines at the N-terminal region of the core histones H2A, H2B, H3, and H4 in the nucleus. The previous studies described the mechanism of interaction and regulation of HDAC4 and HDAC5 by YWAHG. In the inactive state, phosphorylated deacetylases are located in the cytoplasm, where they bind to the 14-3-3 domain of YWHAG via three phosphorylated sites. These interactions lead to the sequestration of HDAC4/5 to the cytoplasm (**Grozinger et al. 2000**; **M. J. Wang K. et al. 2000**). This keeps HDAC4 from entering the nucleus and repressing the transcription of genes important for different functions like neuron development (**M.-S. Kim et al. 2012**; **Pennington et al. 2018**). YWHAG is linked to a type of developmental and epileptic encephalopathy that is characterized by neurodevelopmental impairment and the onset of seizures leading to delays in cognitive and motor development, whereas mutations in HDAC4 are found in patients with neurodevelopmental disorder with central hypotonia and dysmorphic facies (NEDSHF), brachydactyly and intellectual disability. To illustrate how understanding the interaction mechanism can be informative about the variant impact and the potential cause of the disease, consider the Glu247Gly mutation within the RKTASEP motif in HDAC4 is associated with NEDSHF (**Wakeling et al. 2021**). This mutation is documented as a pathogenic missense variant in the ClinVar database. It is not reported in gnomAD and has been determined as a de novo mutation. It was functionally studied, where immunoprecipitation with HDAC4 with the Glu247Gly mutation

in HEK293 cells demonstrated a reduced binding affinity for another 14-3-3 protein, YWHAB (**Wakeling et al. 2021**). As the PPI interface is the same as with the YWHAG protein, we can assume this mutation might also disrupt the interaction with the 14-3-3 domain like YWHAG. By knowing the mechanism of interaction we can hypothesize that the resulting reduced binding or loss of interaction with YWHAG may lead to the impaired nuclear export of HDAC4, causing abnormal expression of genes and contributing to the disorder.

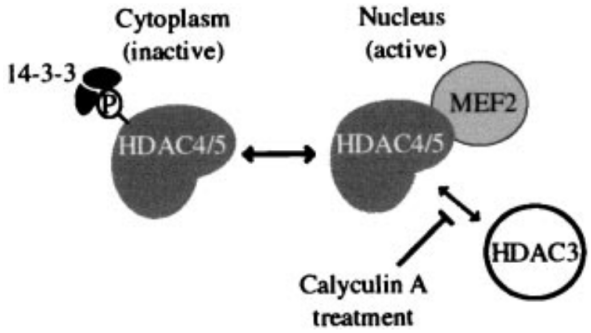


Figure 1.9: Model of activity of HDAC4 through the interaction with 14-3-3 domain protein.
Upon phosphorylation of HDAC4, the phosphorylated ligand motif is recognized by the 14-3-3 domain. This domain-motif interaction leads to the sequestration of HDAC4 and HDAC5 to the cytoplasm, preventing them from downregulating gene transcription (Grozinger et al., 2000).

1.4 Predicting the known occurrence of DMIs in protein interactions using sequence-based approaches

The most efficient way to characterize DMIs would involve sequence-based analysis and structural modeling. This combined approach includes two steps: 1) using sequence-based predictions to identify potential contact residues between proteins, and 2) structural modeling to visualize and pinpoint inter-atomic interactions at the interface. Furthermore, the predicted structural model of the putative interface can aid in the experimental validation by designing the mutations assumed to perturb the binding between the interacting regions. I will discuss this part in more detail in Section 2, Article II.

One way to predict DMIs is by identifying the instances of known DMI types. Databases like ELM contain a catalog of high-quality DMI

types manually curated based on experimental evidence. As ELM employs the regular expression patterns (**see section 1.2.3**) and HMMs of the corresponding binding domains, it can help find known occurrences of similar domains and motifs in the protein interactome (**Weatheritt et al. 2012**; **Edwards et al. 2014**; **Gouw et al. 2018**).

The interaction of Eukaryotic Linear Motif (iELM) is the web server that employs the annotated motifs from ELM and PPI data to identify putative SLiM-mediated interactions extracted from the STRING database (**Weatheritt et al. 2012**). The iELM first checks for domain-domain interfaces using the 3did, the DDI database (**Mosca et al. 2014**). If DDI is found, then the search stops. If no such interface is found, it predicts motifs by employing ELM resource regular expressions and aligning the sequence of the queried protein with their orthologs. Predictions are scored using the SLiMSearch algorithm based on motif conservation (**Davey et al. 2011**). Next, putative motifs and the flanking regions are evaluated for the intrinsic disorder propensity by the IUPred tool (**Dosztányi et al. 2005**). Concurrently, motif-binding domains are detected via the HMMSearch and optionally using Pfam HMMs (**J. Finn M. et al. 2010**). The E-value derived from the HMM match, conservation, and disorder score of identified motifs is used to train a support vector machine to evaluate putative DMIs. If templates for the putative DMIs are available, structural modeling is performed by PepSite, which scores the biophysical feasibility of modeled DMIs (**Petsalaki et al. 2009**). The benchmarking iELM achieved a sensitivity of 84.8 % and a specificity of 86.5 % on its test set (**Weatheritt et al. 2012**).

Despite its good performance, the evaluation of iELM was done on the imbalanced dataset, where the number of negative points outnumbers the positive data points by almost 30-fold. Also, iELM halts the search of potential DMIs, if any domain-domain interface type is found. Since DMIs and DDIs are not mutually exclusive and can act synergistically in interactions this approach may overlook potential DMIs. Moreover, iELM builds HMMs tailored to specific motif-binding domains using hand-curated sets of known sequences. This approach carries the risk of overfitting, as HMMs can become too specialized for a narrow domain data set. Additionally, iELM was not updated and is no longer in use. These limitations motivated my former colleague to develop a DMI predictor tool, that I applied and experimentally validated putative DMIs. The workflow of the tool and its application will be covered in Chapter 3.

While DMI interface predictions can be made, systematic experimental validation has to be done. Below, I will discuss various large-scale

methods and suggest suitable assay for the proposed strategy.

1.5 Systematic experimental validation of putative DMIs

Today, various high-throughput methods for the systematic discovery of PPIs have been developed.

Validation of putative interfaces can be done by using PPI interaction assays that quantify the effects of mutations on PPIs, where mutations, for example, were designed to validate predicted interfaces or were found in patients. When mutations, designed to validate predicted interfaces or identified in patients, reduce or eliminate binding compared to the wild-type, it suggests that the interface is involved in the interaction.

However, the disruptive effect on the interaction by mutation can be caused by other reasons such as partial misfolding, or complete unfolding leading to the destabilization of the protein or its degradation. Alternatively, it can cause the mislocalization of the other subcellular compartment and/or further lead to protein degradation. Therefore, it is essential to use a method that allows monitoring of protein expression levels and provides a quantitative score indicating the binding strength of interactions.

In this section, I describe different *in vitro* and cell-based methods capable of identifying PPIs, and potential assays suitable for experimental validation of putative DMIs.

PPI methods are broadly classified into binary methods or co-complex methods **(Table 1, 1-2)**. For example, AP-MS is known for its scalability in the systematic interaction mapping (**see section 1.1**). Due to the design and principle of the method to detect protein associations rather than direct PPIs, it would not be effective for domain-interface validation. Moreover, this assay may fail to detect transient or weak interactions during lysis and washing steps.

On the other hand, in-vitro methods like ITC, SPR, FP and MST (**see Table 1, 3-6**) detect likely direct interactions and provide real-time information on the binding affinity of these PPIs (**Ward2001; Stahelin2013**; **Pierce et al. 1999**). While these methods are quantitative and can assess the effect of mutations on interactions, they require purified proteins, which can be time-consuming and expensive equipment making these assays less scalable. Due to complications in the purification step, only potentially binding protein fragments are used, making it unclear how the interaction occurs in a full-length context. Additionally, since these assays operate outside the native cel-

lular context, the validation of domain-motif interactions (DMIs) in cells remains uncertain.

Another method is Cross-linking (XL-MS), performed in both *in vitro* and in cell-based systems (**see Table 1, 7**) is valuable for discovering new interfaces, as it captures contact residues in close proximity and provides structural insights. However, it is less suited for interface validation. For instance, the washing step may fail to catch DMI-driven PPIs and inefficient cross-linkers may capture intra-protein contacts, complicating the analysis. Designing mutations for validation can be challenging, as the cross-linkers target specific residues. This method does not allow for comparing the effect of mutation on the binding affinity of PPIs compared to the wild-type proteins. While useful for discovering new interfaces, XL-MS is not suitable for validating interaction interfaces.

Assay Name	Type	Assay is based on…	Assay detects…	Scalable?	Able to test potential effect of mutation on specific PPI?	Able to study the effect of mutation on binding affinity of specific PPI?	Able to measure protein expression levels?	Able to check protein localization?	Comments
Affinity Purification (AP)-Mass Spectrometry (MS)	In vitro	Affinity purification* of a bait protein along with its prey (associated) partners, followed by MS	Protein complexes	Yes	No	No	No	No	*protein purification might be time-consuming and large sample amounts are required
Co-immunoprecipitation (Co-IP) -Mass Spectrometry (MS)	In vitro	The use of specific antibodies* to pull down a target protein along with its prey (associated) partners,, followed by MS	Protein complexes	No	No	No	No	No	* Expensive (e.g. due to the need for specific antibodies)
Isothermal Titration Calorimetry (ITC)	In vitro	Measuring heat changes if two proteins interact	Likely direct PPI	No	Yes	Yes	No	No	
Surface Plasmon Resonance (SPR)	In vitro	Measuring changes in refractive index to quantify binding	Likely direct PPI	Yes	Yes	Yes	No	No	
Microscale Thermophoresis (MST)	In vitro	Measuring the thermophoretic movement of molecules in a temperature gradient to quantify binding.	Likely direct PPI	No	Yes	Yes	No	No	
Fluorescence Polarization (FP)	In vitro	Measuring changes in the polarization of fluorescent light emitted by a fluorophore.	Likely direct PPI	No	Yes	Yes	No	No	
Cross-linking Mass Spectrometry (XL-MS)	In vitro / Cell-based	Using chemical cross-linkers to capture protein-protein interactions, followed by mass spectrometry to identify cross-linked peptides.	Likely direct PPI	Yes	No*	No	No**	No**	*Not suitable for this (or the mutation design is quite complicated, as cross-linkers recognise specific residues, therefore the mutation has to be done or occur on them) **Can be if it is cell-based assay, where proteins are tagged followed by measurement of the tag signal (e.g. fluorescence)
Yeast Two-Hybrid (Y2H)	Cell-based*	DNA-binding and activation domains fused to interacting proteins	Likely direct PPI	Yes	Yes	No	No**	No**	*Proteins are forced to be in the nucleus of the yeast **If the proteins are tagged prior to transformation and checked by flow cytometry and microscopy
Protein Fragment Complementation Assay (PCA)	In vitro / Cell-based	The reconstitution of a transcriptional activator when two proteins of interest interact	Likely direct PPI	Yes	Yes	Yes	No	No	
Proximity Ligation Assay (PLA)	Cell-based	using proximity-dependent ligation of oligonucleotide-conjugated antibodies to create a signal that is amplified and quantified if the proteins are within close proximity.	Likely direct PPI	No	Yes	No	No	No	
Fluorescence Resonance Energy Transfer (FRET)	In vitro / Cell-based	Measures energy transfer between two fluorophores	Likely direct PPI	Yes	Yes	Yes	Yes	No*	*The localisation can be checked if combined with imaging
Bioluminescence Resonance Energy Transfer (BRET)	In vitro / Cell-based	Detecting the energy transfer between a bioluminescent donor and a fluorescent acceptor when they are in close proximity.	Likely direct PPI	Yes	Yes	Yes	Yes	No*	*The localisation can be checked if combined with imaging
MAPPIT (Mammalian Protein-Protein Interaction Trap)	Cell-based	Reconstituting the JAK/STAT signaling pathway through the interaction of bait and prey proteins, leading to reporter gene activation	Likely direct PPI	Yes	Yes	Yes	No	No	
Luminescence-based two-hybrid assay (LuTHy)	Cell-based followed by Cell-free	BRET based assay followed by Co-IP	Likely direct PPI	Yes	Yes	Yes	Yes	No*	*The localisation can be checked if combined with imaging

Table 1 The overview of different *in vitro* and cell-based methods to detect PPIs.

In parallel, cell-based methods have been developed. Cell-based binary methods detect PPIs mostly based on co-expression of genetically tagged proteins. If these proteins interact, their tags come into proximity, producing various readouts to indicate a PPI. For example, common read-outs include the reconstitution, activation or expression of reporter proteins. A well-known example is the Yeast Two-Hybrid (Y2H) assay, where the DNA-binding domain is fused to a bait protein, and the transcription activation domain is fused to a prey protein (**Chien et al. 1991**; **Fields et al. 1989**). When the bait and prey interact, the transcription factor is reconstituted, activating the reporter gene. The presence of interaction is indicated by the activation of the reporter gene and the growth of the yeast.

While Y2H has been attempted to be used for interaction profiling and studying the effects of mutations on PPIs (**see section 1.1**), it cannot directly indicate whether reduced yeast growth is due to a partial misfolding, or unfolding of the proteins, as Y2H does not allow to monitor the protein expression in a real time. Other additional techniques like western blotting are needed for validation. On the other side, fluorescent tagging of proteins before transformation followed by flow cytometry can also be used to check protein levels, but this requires additional steps, costs and expertise. Overall, while Y2H is a simple and useful assay for detecting PPIs, its limitations hinder its ability to fully characterize interaction interfaces.

Following principles similar to Y2H, many binary methods have been subsequently developed to mitigate the shortcomings of Y2H. Examples of these methods include the Protein Fragment Complementation Assay (PCA) and the Mammalian Protein-protein Interaction Trap (MAPPIT) assay (**see Table 1, 9-10**). PCA, a reporter protein (e.g. GFP) is split into two non-functional fragments. These fragments are genetically fused to the proteins of interest, one to each fragment. When the two proteins interact, the fragments come into proximity, allowing the reporter protein to reassemble and regain its functional state, which serves as a readout for the interaction. The advantage of this assay is the detection of likely direct PPIs in living mammalian cells, therefore providing a more optimal cellular context for testing human proteins for the interaction. Similarly, MAPPIT is based on the reconstitution of the JAK-STAT signaling pathway, a key pathway involved in cytokine-mediated signal transduction. In MAPPIT, a mutated cytokine receptor fused to a bait protein recruits JAK upon interaction with a prey protein, leading to STAT activation and reporter gene transcription (**Lievens et al. 2011**). Due to the involvement of this pathway, the assay is limited to PPIs that occur near the plasma membrane. Additionally, steric hin-

drance might interfere with potential interactions. Both methods provide the mammalian context of tested interactions, but it is not possible to monitor protein expression which is essential for the characterization of putative DMIs.

Alternative to the methods that rely on the reconstitution of the reporter protein, methods like Förster resonance energy transfer (FRET) and Bioluminescence resonance energy transfer (BRET) offer more direct readouts based on physical proximity. These assays detect PPIs through non-radiative energy transfer between a donor and acceptor molecule, which occurs only when they are in close proximity. In FRET, proteins are fused with donor and acceptor fluorophores, and upon interaction, energy is transferred from the donor to the acceptor, generating a detectable fluorescent signal (**Sekar et al. 2003**; **S. S. Vogel et al. 2006**; **Grünberg et al. 2013**).

In BRET, luciferase is used as a donor and fluorescent protein acts as an acceptor. The donor is not excited with monochromatic light at its specific excitation wavelength. Instead, the luciferase donor is activated by a chemical substrate, such as coelenterazine-h. This substrate undergoes oxidation by the luciferase enzyme leading to the emission of light (**Xu et al. 1999**). For example, the Nanoluc luciferase tag when using coelenterazine-h emits light with a maximum wavelength of 460 nm (**Hall et al. 2012**). Upon the addition and oxidation of the substrate, when the proteins are in proximity, the energy is transferred from the donor to the acceptor. The emitted luminescence is commonly detected at the short wavelength of the donor, and the long wavelength of the acceptor. The ratio of acceptor energy over donor is the BRET ratio, indicating the potential interaction (**Pfleger et al. 2006**). Both methods provide real-time study of transient PPIs in living mammalian cells. Monitoring protein expression levels is crucial for characterizing interaction interfaces and understanding potential interaction failures. To quantify the binding affinity, saturation experiments can be also performed in both methods, where the quantity of one interaction partner is kept constant while increasing amounts of the other protein (**Pfleger et al. 2006**; **Trepte et al. 2018**).

Along with monitoring the expression levels of proteins, it is possible to check the localization of mutated proteins relative to their wild-types. For example, it can be achieved with bioluminescence imaging (BLI) using high-content screening (HCS) microscopy (**J. Kim et al. 2024**). In a high-content screening system, a plate with the co-expressed tagged proteins in cells is visualized. The HCS is equipped with high-sensitivity cameras and appropriate filters to detect the specific wavelengths of light emitted by the tags. First, the fluorescence expression proteins

are captured and then upon the addition of substrate luminescence is measured.

The main limitation lies in the sensitivity of the assay and the orientation of the tag. As it relies on the proximity it may catch indirect interactions that are involved in a protein complex. In contrast, the real PPI might not be detected due to the steric hindrance of the tags leading to false negatives.

In contrast to FRET, BRET offers several advantages:

- the use of luminescence and the substrate in BRET excludes acceptor cross-excitation and donor photobleaching, which simplifies data analysis
- the reduced auto-fluorescence
- luciferase provides a high sensitivity due to increased signal-to-background ratios
- lower amounts of DNAs are sufficient due to the high sensitivity

Hence, these advantages make BRET a suitable approach for the validation of putative DMI interfaces. In a recent study Wanker and colleagues combined BRET and Co-IP with a luminescence-based readout in one method (**Trepte et al. 2018**). This method named luminescence-based two-hybrid assay, shortly LuTHy, provides a double-readout for PPI detection, which enhances the confidence of identified PPIs in a high-throughput.

Overall, the advantages of the BRET assay might be the optimal choice to be incorporated into the proposed strategy to tackle the questions addressed in my study. Wanker lab kindly provided us with the necessary donor and acceptor vectors, as well as the controls for our study described in Chapters II and III.

1.6 Aims of the thesis

Despite advances in sequencing technologies, most genetic variants remain poorly understood, hindering our grasp of disease mechanisms and complicating clinical diagnosis and treatment (**see subsection 1.1.2**).

Edgotyping has been proposed as a strategy to address this challenge (**see subsection 1.1.2**). While several studies have attempted to apply this strategy using Y2H, to do it entirely experimental is laborious and expensive and, therefore less efficient. To address this question, my goal of the study is to propose a systematic approach enabling the characterization of variant effects by predicting DMIs and using this information

for PPI profiling. This approach will include both computational and experimental methods.

First, I aimed to build the experimental pipeline to validate putative DMI interfaces. To achieve this, I need binary PPI data that can serve as a resource for discovering new interfaces. The HuRI dataset is the largest dataset of binary protein-protein interactions (**Luck et al. 2020**) described in thesis section 1.1.2. In our lab, we have full access to the open reading frame (ORF) HuRI collection. However, the ORFeome collection currently exists in a single copy, while creating multiple copies for use and storage is essential. Since cloning procedures and site-directed mutagenesis are necessary for mutating proteins of interest and testing PPIs I will probe the cloning in tube format first, then adapt it to the plate format. Moreover, as the BRET assay has not yet been established in our lab, I will also assess its sensitivity.

The second aim is to employ a data-driven approach to select PPIs with predicted DMIs suitable for experimental validation. First, DMIs need to be predicted. My colleague developed a DMI tool, used this tool to generate predictions and mapped putative DMIs on the HuRI PPI dataset. To get mutations that may fall into predicted DMIs, another colleague processed the mutations from the largest patient database, ClinVar and overlapped the ClinVar mutations with mapped interface predictions. To further select PPIs mapped with DMIs and overlapped with mutation data suitable for the experimental validation, I need to know which ORFs and which isoforms are available in the ORFeome collection, and how many of those are cloneable and present in a full-length context. Moreover, manual inspection of predicted DMIs for biological relevance will be done. To do these proteins will be annotated with experimental and biological information. Furthermore, for the experimental validation of selected PPIs, controls such as known DMI-mediated PPIs and the PPIs mediated by different interfaces like DDI served as positive and negative controls will be also chosen and included in the study.

Chapter 2

The development of the medium-throughput cloning and the BRET assay pipeline for the experimental validation of predicted DMIs

2.1 Preparation of the wild-type human ORFeome collection

As stated in Aim 1, the availability of a comprehensive ORFeome collection is essential for my project. This collection provides access to GATEWAY-compatible clones for most wild-type proteins from the HuRI dataset, which are necessary for cloning into LuTHy expression vectors that will be used in interaction profiling.

My supervisor Katja Luck brought one copy of a human ORFeome collection from her PostDoc lab comprising ORFs for around 17,500 human protein-coding genes. These ORFs are stored as GATEWAY-compatible clones, allowing them to be transferred to the destination vectors carrying the fluorescence and luminescence reporter tags needed for BRET assay. As this collection only came in one copy, for maintenance and safety reasons, together with my colleague, I adapted and optimized the protocol for making 3 copies of the ORFeome using the Rainin liquidator 96 Manual pipetting system, kindly provided by Khmelinskii group (**Figure 2.1**). The first copy serves as a working collection, the second copy will be backed up and the final copy is supposed to be given to the media lab for the IMB community as an open resource. Overall, the 2-day protocol enables handling 16 96-well plates

for making three copies. Original plates are thawed and fresh plates with media are inoculated and placed for incubation overnight. The incubation can be challenging due to a vast evaporation effect that leads to losing the volume needed to make three copies on the second day after the incubation. Therefore, to optimize this step we tested different incubators, materials to seal plates, boxes to cover plates in the incubator, and testing the well volumes we could use. In two months, we successfully copied 238 plates.

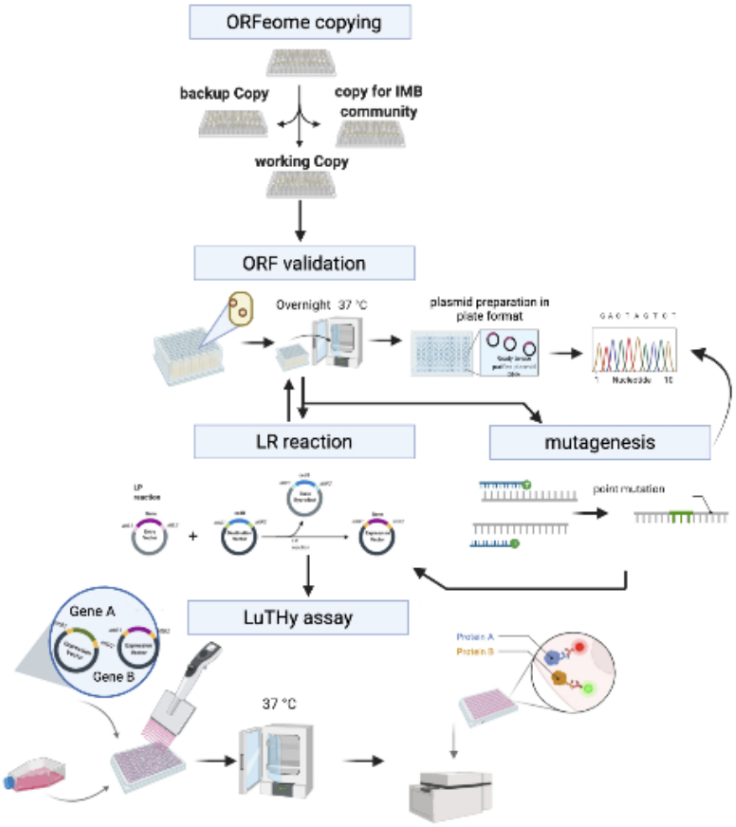


Figure 2.1: The scheme of cloning pipeline followed by BRET assay. The ORFeome collection was copied, where one was a backup copy, the second was made for the IMB community and the third served as a working copy. The ORFs selected for cloning were incubated in 96 deep-well plates overnight, and DNA on the next day was verified by sequencing. Next, the clones were shuffled from the donor GATEWAY vector to the destination vector using the LR reaction. The mutant constructs are generated by site-directed mutagenesis and sequence verified. Upon cloning, the BRET assay is performed.

2.2 The assessment of the sensitivity of BRET assay

To evaluate the sensitivity of the assay chosen for the experimental validation I needed to adapt the cloning of GATEWAY vectors from one tube to plate format and adapt the mutagenesis protocol to a medium-

throughput pipeline. To do this I used the open reading frames (ORFs) coding for proteins, mutations and PPIs as well as controls from my collaborative project with the Koenig (IMB) and Sattler (Institute of Structural Biology, Helmholtz German Research Center for Environmental Health) groups.

The Koenig lab has recently established Far Upstream Element Binding Protein 1 (FUBP1) as a novel regulator in mRNA splicing. Our aim in this project is to aid in the identification of PPIs between FUBP1 and known protein components of the 5' and 3' splice sites as well as of the branchpoint on the mRNA and to delineate the corresponding interaction interfaces. For this project, I generated 64 different constructs using the cloning technique followed by BRET assay (**Figure 2.1**). To test the sensitivity of the assay, I transfected different ratios of ORFs in donor-acceptor constructs (1:10 ng, 1:20 ng, 1:50 ng, 1:100 ng, 1:200 ng) into HEK293 mammalian cells for co-expression.

Along with these pairs, I also included the standard controls. Wanker group, developers of the LuTHy method kindly provided us with standard controls including empty vector controls to rule out background effects from the vector, donor-only (NanoLuc) and acceptor-only (mCitrine) constructs to ensure interactions require both constructs and non-interacting protein pairs to check for false positives. A positive control pair with the known protein-protein interaction BAD-BCL2L1 was included to validate the system's functionality.

Additionally, I used random protein pair controls for each tested pair, consisting of proteins not expected to interact, such as those with different cellular localizations (e.g., nuclear proteins paired with cytoplasmic proteins). As proteins of interest are localized in the nucleus and proteins from protein pairs are found in the cytoplasm, we paired up the tested protein of interest with one protein from the positive pair.

I tested all interactions together with controls and quantified BRET. The corrected BRET (cBRET) ratio is calculated by subtracting either the BRET ratio of controls (donor-only (i.e. NanoLuc) and acceptor-only (i.e. mCitrine) constructs) from the BRET ratio of the studied interaction of interest. Our findings showed that cBRET values for the weak interactions were close to the cBRET values of the random pairs. Based on this information, I learned that a high amount of transfected cDNA might lead to the generation of false-positive data. Therefore, we questioned those findings and evaluated assay specificity by testing the range of different DNA ratios of the previously tested interactions. I discovered that 1:50 ng appears to be a good ratio for the discrimination of significant from non-significant cBRET signals (Figure 2.2).

In summary, copying the ORFeome collection and testing BRET as-

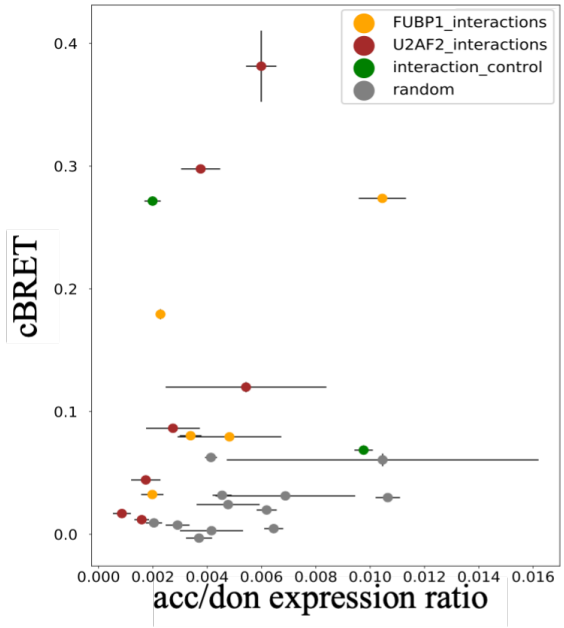


Figure 2.2: The evaluation of BRET's sensitivity with 1:50 ng of donor: acceptor DNA ratio. The plot represents calculated cBRET ratios for tested FUBP1 (orange), U2AF2 (red) interactions, positive controls (green) and random pairs (gray) as a function of acceptor to donor (acc/don) protein expression ratio. All values are the mean +/- s.d. from two technical replicates.

say sensitivity enabled quick access to the ORFs and helped define the ratio of tested constructs needed for the transfection to avoid false positives. This allowed us to explore the application of BRET assay in interaction profiling to further explore protein-protein interactions (PPIs) involved in mRNA splicing.

2.3 Article I: FUBP1 is a general splicing factor facilitating 3' splice site recognition and splicing of long introns

Summary

The splicing of pre-mRNA plays a crucial role in gene regulation and the expansion of the proteome in eukaryotes. However, the information on how the recognition of splice sites and pairing during spliceosome assembly occurs lacks details. This project focused on understanding the role of FUBP1 in RNA splicing, particularly its function in the recognition and processing of 3' splice sites and splicing of long introns. Using *in vivo* iCLIP analysis we found that FUBP1 binds to 91.3 % of 3' splice sites in a similar pattern as core splicing factors like SF1, U2AF2 and SF3B1. Further investigation showed that FUBP1 recognized cis-regulatory RNA motif located upstream of the branch point (BP) in

pre-mRNA. Through EMSA and ITC experiments, we demonstrated that FUBP1 binds GU-rich sequences. This was further validated by NMR and *in vivo* iCLIP data showing that KH domains of FUBP1 independently recognize these motifs. Moreover, kinetic modeling and transcriptional profiling demonstrated that FUBP1 is required for efficient splicing of long introns, which represent 80 % of human introns.

Next, we explored the interactions of FUBP1 with other splicing factors. First, we studied FUBP1 interactions with components of spliceosome complexes. Here, NMR analysis provided insights into the interaction between FUBP1 and U2AF2, the key component 3' splice complex. The preliminary structure from the NMR study suggests that the second RRM domain of U2AF2 and the N-terminal N-box of FUBP1 protein represent the minimal binding regions. Furthermore, we found that the amino acid change from alanine to aspartate at residue 38 (A38D) sitting at the N-box of FUBP1 disrupts the interaction using recombinant fragments of FUBP1 and U2AF2. This data was supported in a full-length context using the BRET experiments in mammalian cells (**Article I, Figure 3, C & J**). Given the obtained BRET data, we observed that the presence of mutation significantly increased the distance between proteins, but the interaction was not completely disrupted. Here, we hypothesize that mutated FUBP1 and U2AF2 still interact due to the binding to the same mRNA but the contact between both is much weaker due to the lost direct interface between both. With BRET we also confirmed the known interaction interface between FUBP1 and SF1, protein of U2 complex at the 5' splice site (**Article I, Figure 3, C & D**). Furthermore, we tested the interactions of FUBP1 with U1 snRNP-associated proteins, including SNRPA, SNRPC, TIAL1, PRPF40B and SNRBP as well as TCERG1 and KHDRBS1. We further confirmed these interactions of FUBP1 with U1-associated proteins with BRET and/or NMR. Interestingly, NMR analysis proposed that FUBP1's A/B boxes interact with proline-rich regions from SNRPB. These changes were less pronounced with SNRPA and PRPF40B containing similar proline-rich stretches (**Article I, Figure 6, G**).

Overall, this study provided a comprehensive analysis demonstrating the global role of FUBP1 pre-mRNA splicing processes. Our key findings suggest that FUBP1 acts as a general splicing regulator at the 3' splice site. Moreover, many tested interactions mediated via domain-motif interface were able to be detected by BRET, and disruption of these interfaces by point mutations established this assay as a valuable system to validate predicted DMIs.

Statement of contribution

This is a collaborative project in which I conducted the following aspects of the study: cloning of the open reading frames (ORFs) and their mutants, assessing the specificity of the BRET assay by finding the optimal DNA ratio for transfection and introducing random pairs (proteins not known to interact with each other) as controls. I also tested wild-type and mutant interactions using BRET and BRET titration experiments. I participated in BRET-driven data analysis and visualization. I also took part in figures and tables preparation for the manuscript and participated in the writing methods part and revision of the manuscript.

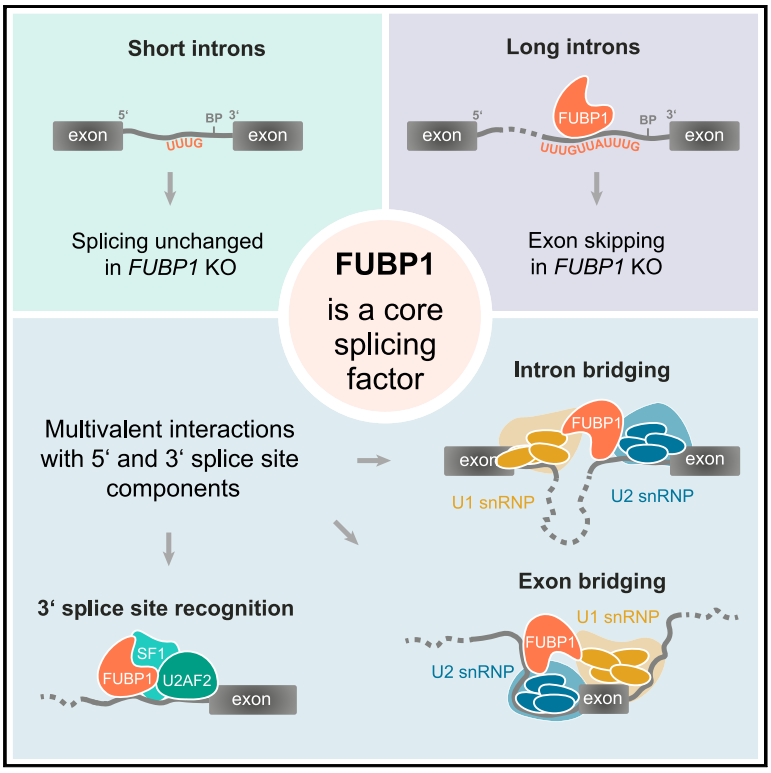
Supervisor confirmation

Molecular Cell

Article

FUBP1 is a general splicing factor facilitating 3′ splice site recognition and splicing of long introns

Graphical abstract



Highlights

- FUBP1 recognizes a ubiquitous *cis*-regulatory RNA motif upstream of the branch point
- Multivalent interactions in disordered FUBP1 regions support spliceosome assembly
- FUBP1 affects long introns, which are prevalent in humans and altered in cancer
- Kinetic modeling and protein interactions implicate FUBP1 in splice site bridging

Authors

Stefanie Ebersberger, Clara Hipp, Miriam M. Mulorz, ..., Katja Luck, Michael Sattler, Julian König

Correspondence

k.luck@imb-mainz.de (K.L.), michael.sattler@helmholtz-munich.de (M.S.), j.koenig@imb-mainz.de (J.K.)

In brief

Ebersberger et al. identify the RNA-binding protein FUBP1 as a key splicing factor that binds to a hitherto unknown *cis*-regulatory motif at 3′ splice sites. Multivalent interactions of FUBP1 with splice site components support spliceosome assembly at multiple stages and ensure efficient splicing of long introns.

Ebersberger et al., 2023, Molecular Cell *83*, 2653–2672
August 3, 2023
https://doi.org/10.1016/j.molcel.2023.07.002

CellPress

Molecular Cell

CellPress
OPEN ACCESS

Article

FUBP1 is a general splicing factor facilitating 3′ splice site recognition and splicing of long introns

Stefanie Ebersberger,[1,12] Clara Hipp,[2,3,12] Miriam M. Mulorz,[1,12] Andreas Buchbender,[1] Dalmira Hubrich,[1] Hyun-Seo Kang,[2,3] Santiago Martínez-Lumbreras,[2,3] Panajot Kristofori,[4] F.X. Reymond Sutandy,[1] Lidia Llacsahuanga Allcca,[1,13] Jonas Schönfeld,[1] Cem Bakisoglu,[5] Anke Busch,[1] Heike Hänel,[1] Kerstin Tretow,[1] Mareen Welzel,[1] Antonella Di Liddo,[1] Martin M. Möckel,[1] Kathi Zarnack,[5,6] Ingo Ebersberger,[7,8,9] Stefan Legewie,[10,11] Katja Luck,[1,*] Michael Sattler,[2,3,*] and Julian König[1,14,*]

[1]Institute of Molecular Biology (IMB) gGmbH, 55128 Mainz, Germany
[2]Institute of Structural Biology, Helmholtz Center Munich, 85764 Neuherberg, Germany
[3]Bavarian NMR Center, Department of Bioscience, School of Natural Sciences, Technical University of Munich, 85747 Garching, Germany
[4]Department of Systems Biology, Institute for Biomedical Genetics (IBMG), University of Stuttgart, 70569 Stuttgart, Germany
[5]Buchmann Institute for Molecular Life Sciences & Institute of Molecular Biosciences, Goethe University Frankfurt, 60438 Frankfurt am Main, Germany
[6]CardioPulmonary Institute (CPI), 35392 Gießen, Germany
[7]Applied Bioinformatics Group, Institute of Cell Biology and Neuroscience, Goethe University Frankfurt, 60438 Frankfurt am Main, Germany
[8]Senckenberg Biodiversity and Climate Research Center (S-BIK-F), 60325 Frankfurt am Main, Germany
[9]LOEWE Center for Translational Biodiversity Genomics (TBG), 60325 Frankfurt am Main, Germany
[10]Department of Systems Biology, Institute for Biomedical Genetics (IBMG), University of Stuttgart, 70569 Stuttgart, Germany
[11]Stuttgart Research Center for Systems Biology (SRCSB), University of Stuttgart, 70569 Stuttgart, Germany
[12]These authors contributed equally
[13]Present address: University of California, Berkeley, CA 94720, USA
[14]Lead contact
*Correspondence: k.luck@imb-mainz.de (K.L.), michael.sattler@helmholtz-munich.de (M.S.), j.koenig@imb-mainz.de (J.K.)
https://doi.org/10.1016/j.molcel.2023.07.002

SUMMARY

Splicing of pre-mRNAs critically contributes to gene regulation and proteome expansion in eukaryotes, but our understanding of the recognition and pairing of splice sites during spliceosome assembly lacks detail. Here, we identify the multidomain RNA-binding protein FUBP1 as a key splicing factor that binds to a hitherto unknown *cis*-regulatory motif. By collecting NMR, structural, and *in vivo* interaction data, we demonstrate that FUBP1 stabilizes U2AF2 and SF1, key components at the 3′ splice site, through multivalent binding interfaces located within its disordered regions. Transcriptional profiling and kinetic modeling reveal that FUBP1 is required for efficient splicing of long introns, which is impaired in cancer patients harboring *FUBP1* mutations. Notably, FUBP1 interacts with numerous U1 snRNP-associated proteins, suggesting a unique role for FUBP1 in splice site bridging for long introns. We propose a compelling model for 3′ splice site recognition of long introns, which represent 80% of all human introns.

INTRODUCTION

Splicing is a crucial step in eukaryotic mRNA processing, and its dysregulation is a hallmark of many cancers.[1–3] Splicing is catalyzed by the spliceosome, a megadalton machinery comprising five small nuclear ribonucleoprotein (snRNP) complexes named U1, U2, U4, U5, and U6.[4–7] During early spliceosome assembly (E complex formation), the 5′ and 3′ splice sites are recognized: U1 binds at the 5′ splice site, whereas U2 auxiliary factor 1 (U2AF1), U2AF2, and splicing factor 1 (SF1) assemble at the 3′ splice site,[6–11] where they specifically recognize AG dinucleotide,[12,13] polypyrimidine (Py) tract,[14–16] and branch point (BP) site, respectively (Figure 1A).[9,17] In the resulting A complex, U2 snRNP is recruited to the BP and stabilized by SF3A and SF3B, and SF1 is released.[18,19] Subsequent snRNP recruitment and further rearrangements (formation of B and C complexes) mediate intron excision and exon ligation to form the mature mRNA.

Strikingly, mechanistic details of splice site recognition by multidomain splicing factors during early spliceosome assembly are lacking.[20,21] U2AF2 binding is central to the early definition of splice sites and is subject to layers of regulation including direct

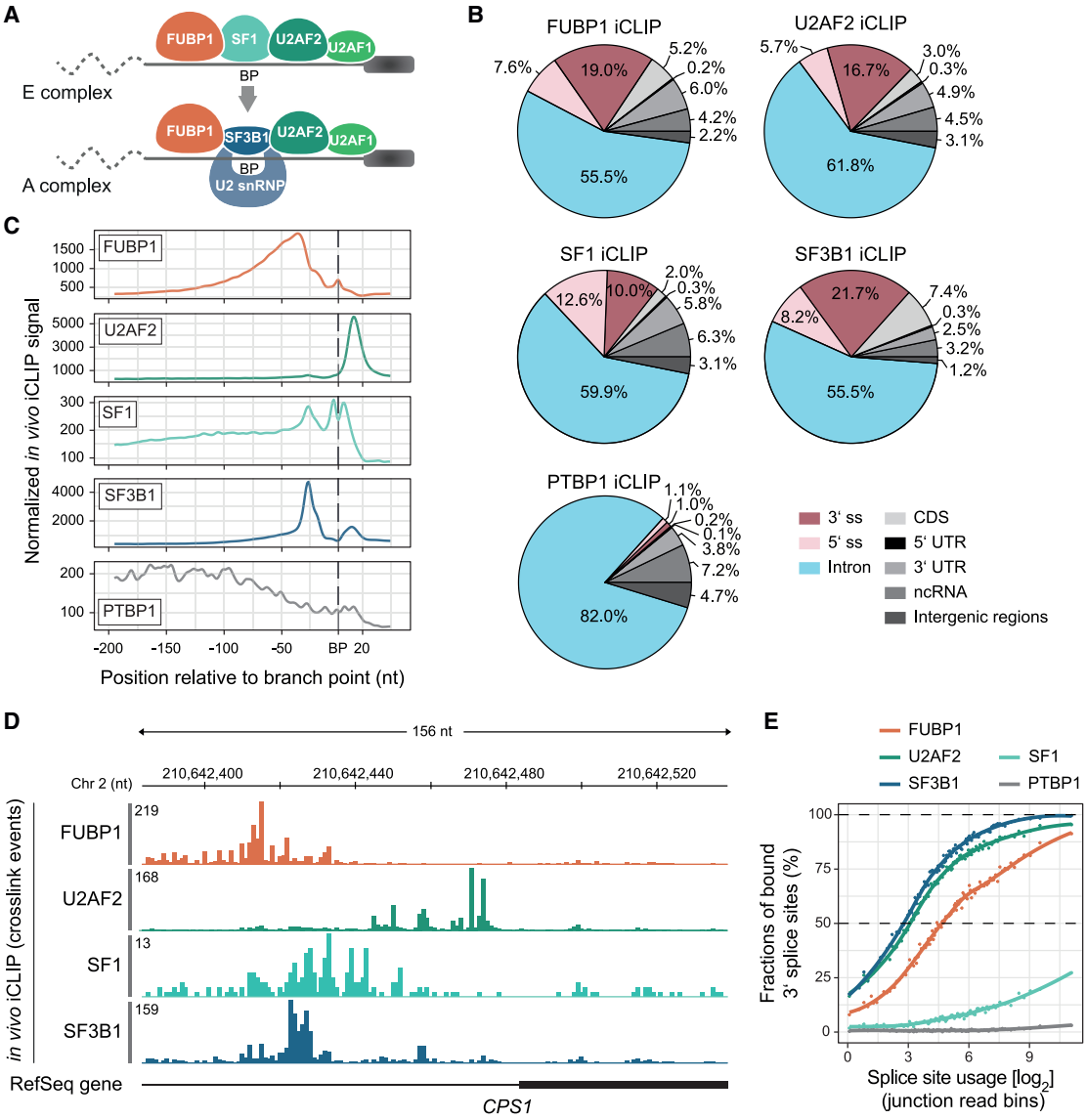


Figure 1. FUBP1 binds upstream of the branch point at 3′ splice sites during early spliceosome assembly *in vivo*

(A) Schematic of spatial RBP assembly at the 3′ splice site in the "commitment" E complex and the pre-spliceosomal A complex. BP, branch point.

(B) iCLIP in HeLa cells. Distribution of binding sites across transcript regions for FUBP1 (n = 854,404), U2AF2 (n = 914,221), SF1 (n = 99,305), SF3B1 (n = 1,694,991), and PTBP1 (n = 127,450). 3′ and 5′ splice sites (ss) refer to 100 nt upstream/downstream of exons, respectively. CDS, coding sequence; UTR, untranslated region.

(C) Metaprofiles of cross-link events of FUBP1, U2AF2, SF1, SF3B1, and PTBP1 relative to the BP.

(D) Genome browser view of an internal exon in the *CPS1* mRNA displaying the iCLIP data for FUBP1, U2AF2, SF1, and SF3B1 from HeLa cells.

(E) Saturation analysis showing the percentage of bound 3′ splice sites for each RBP in each quantile.

competition, cooperative recruitment, change of RNA secondary structure, dynamic conformational states, and autoinhibition.[15,22–30] Despite the pivotal role of U2AF2, the precise contribution of cofactors and multivalent interactions are yet to be elucidated. Recently, we reported how U2AF2 achieves specificity despite the degeneracy of its pyrimidine-rich RNA-binding motif.[28] In this study, we found that the RNA-binding protein (RBP) far upstream binding protein 1 (FUBP1) promotes U2AF2 binding to RNA.

FUBP1 was initially characterized as a transcriptional regulator of the proto-oncogene c-*myc* through binding to AT-rich DNA elements and interaction with PUF60, also known as the FUBP-interacting repressor (FIR).[31–34] However, more recently, FUBP1 has also been reported to bind RNA and to influence translation or splicing of specific transcripts.[35–38] Similar to its DNA-binding specificity, FUBP1 exhibits a general preference for AU- and GU-rich RNA[31] that is expected to derive from its four K homology (KH) domains.[39] Notably, cancer-associated

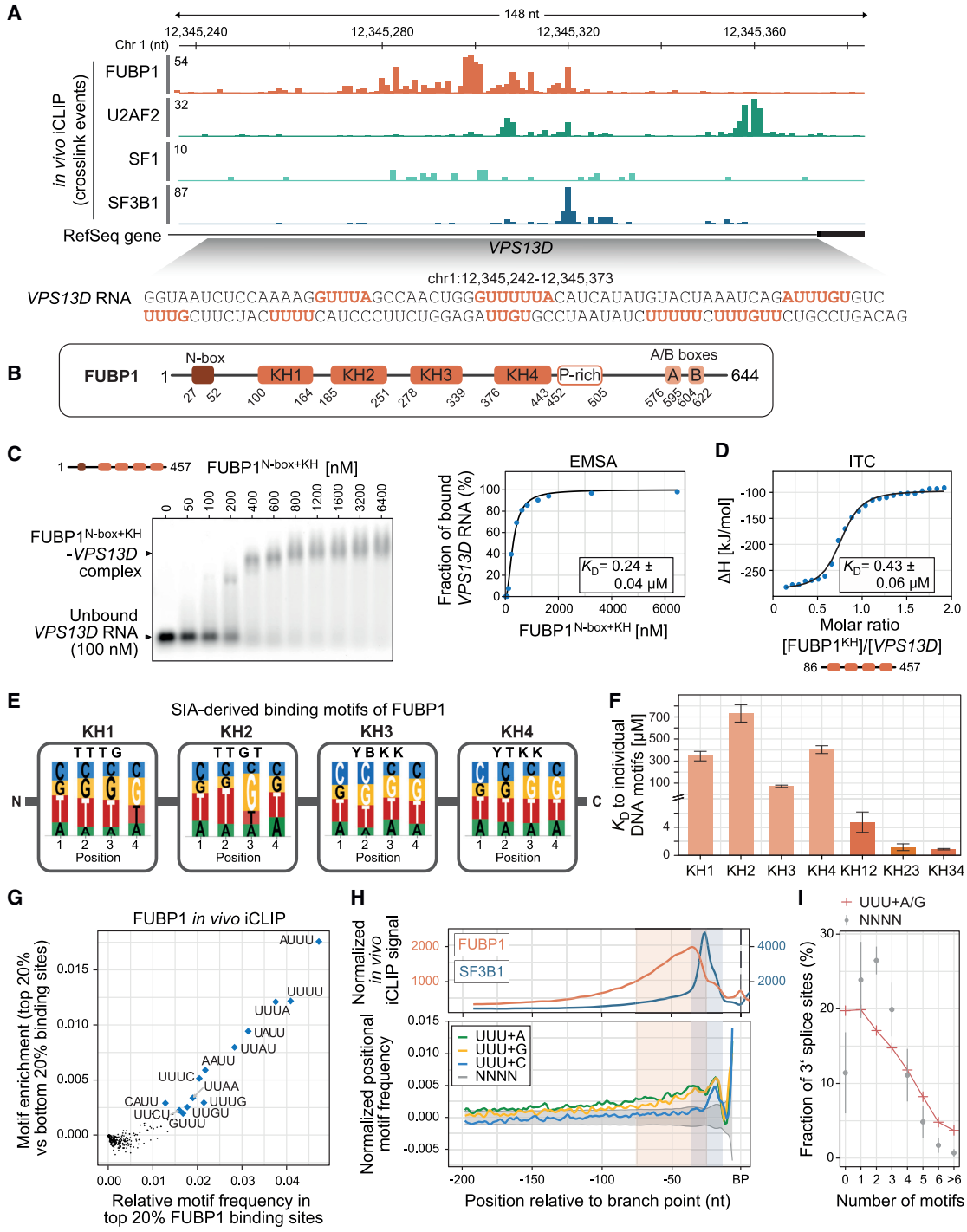


Figure 2. FUBP1 binds a hitherto unknown *cis*-regulatory motif upstream of the BP

(A) Genome browser view of an internal exon in the *VPS13D* mRNA displaying the iCLIP data for FUBP1, U2AF2, SF1, and SF3B1.

(B) Domain architecture of FUBP1. KH, K homology domain; P-rich, proline-rich stretch.

(C) Agarose gel (left) and quantification with fitted curve (right) from an EMSA experiment with recombinant FUBP1$^{N\text{-box+KH}}$ (50–6,400 nM) and a fluorescently labeled 132-nt RNA fragment of *VPS13D* (100 nM). Measurements were performed in duplicates and data are represented as mean ± standard deviation (SD).

(D) Binding affinity for the interaction of FUBP1^{KH} with *VPS13D* RNA determined by ITC. ITC measurements were performed in triplicates and data are represented as mean ± SD.

(legend continued on next page)

loss-of-function mutations within *FUBP1* have been connected to global splicing changes in low-grade glioma,[1,40–42] suggesting an RNA-regulatory role in these processes. Here, we reveal a global role for FUBP1 in pre-mRNA splicing. Our results suggest that FUBP1 functions as a general splicing factor at the 3′ splice site, with a crucial role in promoting efficient splicing of long introns, which make up over 80% of human pre-mRNA transcripts.

RESULTS

FUBP1 is a core component of 3′ splice site recognition

To dissect the role of FUBP1 in splicing, we examined the footprint of FUBP1 and other splicing factors on pre-mRNA in HeLa cells using *in vivo* individual-nucleotide resolution UV cross-linking and immunoprecipitation (*in vivo* iCLIP; Figures 1B and S1A; Table S1).[43,44] As expected, large proportions of the binding sites of SF1, U2AF2, and SF3B1 are located at 3′ splice sites (10%, 17%, and 22%, respectively). Interestingly, FUBP1 shows a similar preference for 3′ splice sites (19%). By contrast, for the more restricted splicing regulator PTBP1, which is known to act on a subset of exons, only 1% of binding sites are located at 3′ splice sites. We confirmed that U2AF2 binds at the Py tract located between the BP and 3′ splice site,[45,46] whereas SF1 binding peaks at the BP, with a reduced signal at the BP adenine itself,[9,17] presumably owing to the lower cross-linking efficiency of adenine (Figures 1C and 1D).[47] Consistent with a previous report,[48] SF3B1 binds in a clamp-wise manner up- and downstream of the BP. Strikingly, FUBP1 also shows a pronounced footprint at the BP (Figures 1C and 1D). Its binding peaks at a location 34 nucleotides (nt) upstream of the BP and tails for up to 100 nt. In comparison, PTBP1 does not display such a ubiquitous positioning at 3′ splice sites (Figure 1C).[49,50] Next, we addressed what fraction of 3′ splice sites is bound using a saturation-based analysis that controls for splice site usage and transcript abundance.[51] We found that FUBP1 binds the same percentage of 3′ splice sites as U2AF2 and SF3B1, which are both universally present at 3′ splice sites (91.3%, 95.4%, and 99.6%, respectively; Figure 1E). By contrast, SF1 and PTBP1 are associated with 27.3% and 3.1% of 3′ splice sites, respectively (Figures 1E and S1B). Overall, these data suggest that FUBP1 functions as a general splicing factor in early spliceosome assembly.

FUBP1 binds a *cis*-regulatory RNA motif upstream of the branch point

Given the prevalence of FUBP1 upstream of the BP, we investigated its RNA-binding preferences. First, we performed electrophoretic mobility shift assays (EMSAs) with a 132-nt RNA fragment upstream of the prototypical 3′ splice site of exon 43 of the *VPS13D* mRNA (*VPS13D*) and a shortened fragment (36 nt) with the region showing the most FUBP1 binding in iCLIP (*VPS13D*^{short}; Figure 2A). We observed strong binding of FUBP1 (FUBP1$^{N\text{-}box+KH}$, aa 1–457) to both RNAs in the low nanomolar range (Figures 2B, 2C, and S1C). Isothermal titration calorimetry (ITC) with *VPS13D* yielded a similar result (Figure 2D; Table S2), confirming the high-affinity binding at this region.

FUBP1 harbors four KH domains, which are expected to bind single-stranded RNA and DNA[32,52] and can act either independently or synergistically[53–55] to recognize extended regions of pre-mRNA. We used nuclear magnetic resonance (NMR) spectroscopy to investigate the modular arrangement of the four FUBP1 KH domains. Superimposition showed that the NMR spectrum of FUBP1^{KH} (aa 86–457) containing KH1–4 was virtually identical to those of the individual KH domains, indicating that the KH domains are structurally independent (Figure S1D). Furthermore, NMR secondary structure analysis revealed that FUBP1 contains KH domains with a typical type I fold that are connected by flexible linkers (Figure S1E).[56] We conclude that the KH domains of FUBP1 are not preformed into an RNA-binding platform but rather can be considered like beads on a string.

To characterize the individual RNA-binding preferences of the four KH domains, we performed a scaffold-independent analysis (SIA), which is based on changes in NMR chemical shifts upon titration with short oligonucleotide motifs (Figure S2A).[57] Initial binding experiments were performed using randomized pools of 5-mer DNA, followed by verification of the identified motifs using RNA oligonucleotides (Figure S2B). SIA identified well-defined consensus motifs for KH1 (UUUG) and KH2 (UUGU) and more loosely defined motifs for KH3 (YBKK, where Y = C or U; B = C, G, or U; K = G or U) and KH4 (YUKK). Hence, all four KH domains exhibit a preference for GU-rich sequences (Figure 2E). The affinities of the individual KH domains to the final motifs, as determined by NMR spectroscopy, are in the high micromolar range (Figures S2C–S2F). Combinations of two KH domains and motifs show strong binding avidity: the ITC-measured affinities for tandem domains were in the high nanomolar to low micromolar range (Figures S2G–S2I; Table S2). This suggests that specificity and high affinity are achieved by avidity and multivalent interactions between the four KH domains and RNA with multiple binding motifs (Figure 2F). Indeed, EMSA and ITC experiments confirmed that multiple FUBP1 binding motifs in the *VPS13D* mRNA fragment increase FUBP1 binding to nanomolar affinity (Figures 2A, 2C, and 2D).

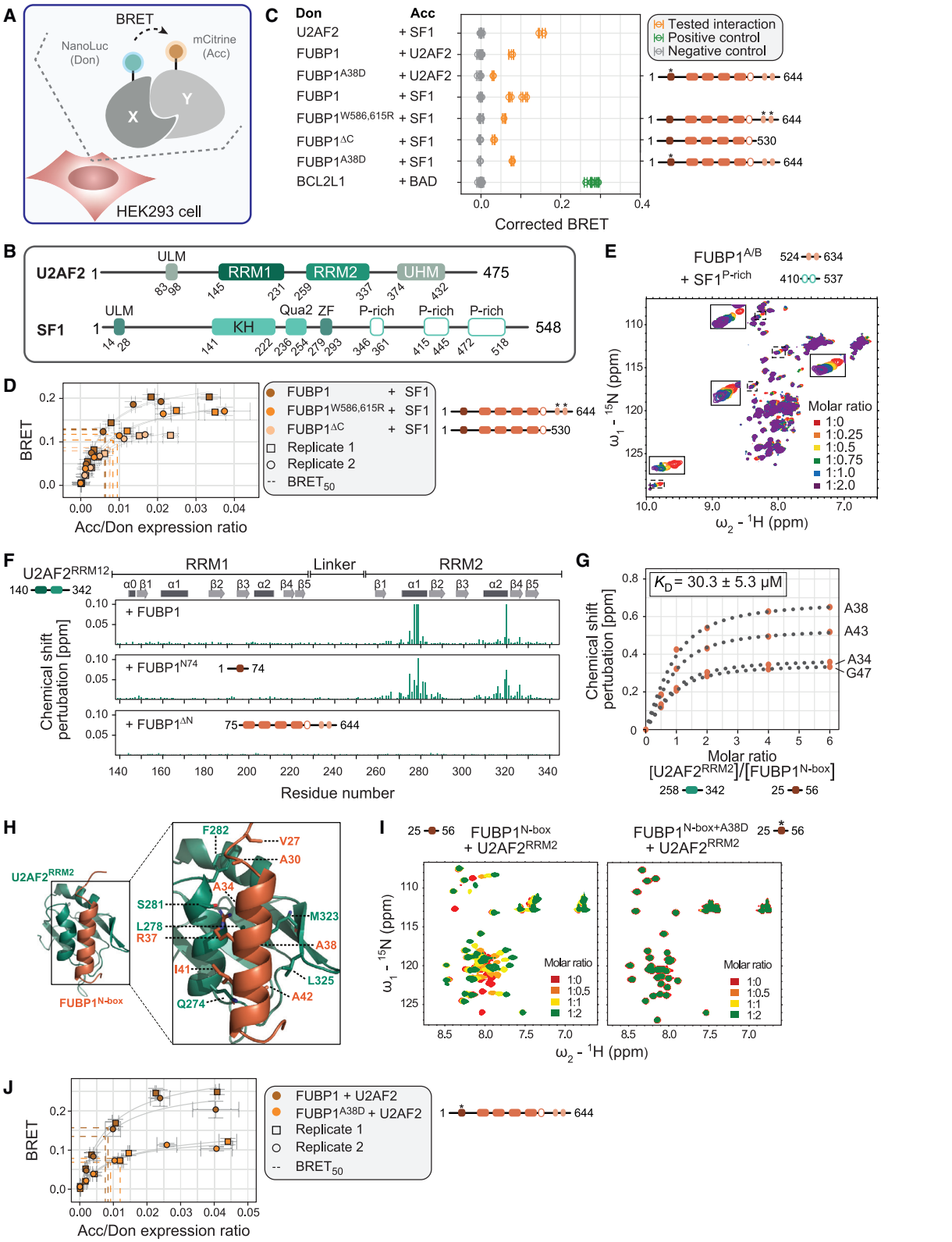
(E) Scaffold-independent analysis (SIA)-derived binding motifs for individual FUBP1 KH domains. Preferred bases are highlighted in white. Y, pyrimidine (T or C); B, not A (C, G, or T); K, keto (G or T).

(F) K_D values of individual and tandem KH domains with their optimal DNA target (KH1, TTTTG; KH2, TTTGT; KH3, TCTGT; KH4, TTTTG; KH1-2, TTTGTAAAATTTTG; KH2-3, TCTGTAAAATTTGT; KH3-4, TTTTGAAAATCTGT) determined by NMR or ITC, respectively (Figures S2C–S2I; Table S2). ITC measurements were performed in triplicates. For NMR, the K_D values of eight selected residues were calculated. Data are represented as mean ± SD.

(G) Motif enrichment in the *in vivo* FUBP1 iCLIP data. Disjunct 4-mer frequencies were calculated for the top vs bottom 20% of binding sites based on expression-normalized iCLIP signals.

(H) Positional enrichment of FUBP1 binding motifs and control motifs relative to the BP. UUU+A/G/C, i.e., 4-mers containing UUU interspersed at any position with A/G/C. NNNN, 100,000 sets of random combinations of four 4-mers. 4-mer frequencies were calculated position-wise upstream of the BP and compared with the average 4-mer frequencies in an intronic control region. Top: Metaprofile of normalized FUBP1 and SF3B1 iCLIP cross-link events at the same 3′ splice sites is shown for comparison.

(I) Abundance of FUBP1 binding motifs at 3′ splice sites of human introns. Background distribution for all possible 4-mers (mean ± 1 SD) is shown in gray.



(legend on next page)

Interrupting the U-rich motifs of *VPS13D*^{short} with cytidines severely reduces the binding affinity, underlining the specificity of the FUBP1-RNA interaction (Figure S1C).

To validate the interaction between FUBP1 and RNA motifs in cells, we compared 4-mer motifs in the sites of strongest FUBP1 binding over background in the *in vivo* iCLIP data. In line with the SIA, we found a strong preference for uridine-rich motifs at FUBP1 binding sites (Figures 2G and S2J). For *in vivo* binding, these motifs can be interspersed at any position by adenine or, to a lesser extent, by guanine. Consistent with the omnipresence of FUBP1 at 3′ splice sites, we observed a striking enrichment of FUBP1 binding motifs (“UUU+A” and “UUU+G,” i.e., three uridines interspersed at any position with adenine or guanine) upstream of the BP, where they coincide with FUBP1 binding (Figure 2H). Conversely, both “UUU+C,” accounting for general uridine richness, and random motif sets are enriched closer to the 3′ splice site but not in the main region of FUBP1 binding. Importantly, enriched FUBP1 motifs upstream of the BP are a common feature across all annotated introns (Figures 2H, 2I, S2K, and S2L), indicating that we identified a previously unknown *cis*-regulatory RNA motif in splicing regulation.

FUBP1 directly interacts with U2AF2 and SF1

Given the prevalence of FUBP1 at functional 3′ splice sites, we examined whether FUBP1 interacts with key early 3′ splice site components in cells using bioluminescence resonance energy transfer (BRET) (Figure 3A).[58] Interaction signals in the BRET assay are indicative of direct contacts or close proximities. As a proof-of-concept, we confirmed the known U2AF2-SF1 interaction.[8,10,18,58] Importantly, we also observed interactions of FUBP1 with U2AF2 and SF1 (Figures 3B, 3C, and S2M), suggesting that FUBP1 is in close or direct contact with these core splicing factors inside cells.

To investigate whether FUBP1 directly interacts with SF1 (Figure 3C), we focused on the C-terminal region of FUBP1, which harbors the A and B boxes (A/B boxes). These motifs are specific to the FUBP family of proteins and have been shown to mediate binding to a proline-rich region of snRNP-U1-70K in fruit flies.[60,59] Similar proline-rich regions are also present in the C-terminal region of SF1. Consistently, the SF1-FUBP1 interaction detected by BRET is reduced upon deletion or mutation of the A/B boxes (Figures 3B–3D and S2M). In addition, ^{1}H-^{15}N correlation NMR spectra of the FUBP1 A/B box region show specific chemical shift perturbations (CSPs) upon titration with the proline-rich region of SF1, indicating direct binding (Figure 3E).

To map the interacting regions in FUBP1 and U2AF2, we performed NMR titration experiments using ^{15}N-labeled U2AF2^{RRM12} [14,15,30] and unlabeled full-length FUBP1. Large CSPs and line broadening in the ^{1}H-^{15}N correlation spectra exclusively map to the U2AF2 RRM2 domain, especially to the two α helices on the backside of the β sheets that mediate RNA binding (Figures 3B, 3F, S2N, and S3A). Moreover, a construct comprising the N-terminal region of FUBP1 (FUBP1^{N74}, aa 1–74) recapitulates the CSPs observed with full-length FUBP1, whereas a construct lacking the N-terminal region (FUBP1$^{\Delta N}$, aa 75–644) does not yield any evident CSPs (Figures 3F and S3A). Complementary NMR titrations with ^{15}N-labeled FUBP1 constructs identify the U2AF2 RRM2 domain and a short peptide motif in the N-terminal region of FUBP1 (aa 27–52), referred to as N-box, as the minimal binding regions (Figures S3B–S3F). The U2AF2 RRM2-FUBP1 N-box interaction exhibits micromolar affinity by NMR titrations (Figures 3G, S3D, S3G, and S3H; Table S2).

To provide a high-resolution view, we determined the NMR-derived solution structure of the U2AF2 RRM2-FUBP1 N-box complex (Figures 3H and S3I–S3K; Table 1). This structure shows a well-defined U2AF2 RRM2 domain and a more mobile helical FUBP1 N-box and reveals that the FUBP1$^{N\text{-box}}$ forms an α helix, which is recognized by helices α1 and α2 and the β4 strand of U2AF2^{RRM2}. Hydrophobic interactions dominate at this interface, where four alanines in FUBP1$^{N\text{-box}}$ (A30, A34, A38, and A42) are aligned along the extended hydrophobic interface, with A38 positioned centrally. Additional contacts involving bulkier side chains, that is, R37 and I41 in FUBP1 and L278 and M323 in U2AF2, further stabilize the binding interface. The recognition of the FUBP1 N-box resembles the interaction between FUBP1 N-box and PUF60,[34] consistent with structural similarities between PUF60 and U2AF2 RRM2 (Figure S4A).[34]

Figure 3. FUBP1 directly interacts with SF1 and U2AF2 via its C-terminal A/B boxes and N-terminal N-box

(A) Schematic of BRET assay. Energy transfer between the substrate oxidized by NanoLuc luciferase (donor, Don) and mCitrine (acceptor, Acc) occurs if proteins X and Y interact.

(B) Domain architecture of U2AF2 (UniProt: P26368) and SF1 (UniProt: Q15637). ULM, U2AF ligand motif; RRM, RNA-recognition motif; UHM, U2AF homology motif family; Qua2, quaking homology 2 domain; ZF, zinc finger.

(C) BRET values for tested interaction pairs and controls. Two biological replicates are shown. Error bars represent SD of technical triplicates. Trp-to-Arg mutations in the A/B boxes were rationalized based on disrupting the hydrophobic contacts as previously reported.[59]

(D) BRET saturation curves for combinations of FUBP1 variants and wild-type SF1. Trp-to-Arg mutations in the A/B boxes or their deletion significantly lowered the maximal BRET signal, although changes in the BRET$_{50}$ (acceptor/donor ratio at which half-maximal BRET signal is reached) were not significant. Amounts of acceptor and donor proteins were estimated by fluorescence and total luminescence, respectively, in intact cells. Two biological replicates are shown. Error bars represent SD of technical triplicates.

(E) NMR titration of FUBP1$^{A/B}$ with SF1$^{P\text{-rich}}$. Significant chemical shift changes are highlighted by boxes.

(F) Binding interface mapping based on NMR titration of U2AF2^{RRM12} with full-length FUBP1, FUBP1^{N74}, and FUBP1$^{\Delta N}$ (Figure S3A).

(G) Binding affinity for the interaction of FUBP1$^{N\text{-box}}$ and U2AF2^{RRM2} from NMR titrations. Chemical shift differences of four exemplary residues of FUBP1$^{N\text{-box}}$ (Figures S3D and S3G) are fitted to binding isotherm to estimate the K_D. Data are represented as mean ± SD of calculated K_D values of eight selected residues.

(H) NMR-derived structure of the complex of U2AF2^{RRM2} (green) and FUBP1$^{N\text{-box}}$ (brown) (Figure S3K; Table 1, PDB: 8P25).

(I) Comparison of NMR titrations of FUBP1$^{N\text{-box}}$ WT and mutant FUBP1$^{N\text{-box+A38D}}$ with U2AF2^{RRM2}.

(J) BRET saturation curves for wild-type FUBP1 and mutant FUBP1^{A38D} against U2AF2. Two biological replicates are shown. Error bars represent SD of technical triplicates.

Table 1. Statistics for structure calculation of the U2AF2^{RRM2}/FUBP1$^{N\text{-}box}$ chimera, related to Figures 3H and S3K, PDB: 8P25[a]

Experimental restraints	
Distance restraints	
Total NOE	2,147
Short range, |i–j| ≤ 1	1,047
Medium range, 1 < |i–j| < 5	392
Long range, |i–j| ≥ 5	708
Dihedral angle restraints (from TALOS)	
Φ	82
Ψ	86
Structure statistics	
RMSD from experimental restraints (mean and SD)	
Distance restraints (Å), no violation > 0.5 Å	0.013 ± 0.007
Dihedral angle restraints (°, no violation > 0.5°)	0.19 ± 0.04
Deviations from idealized geometry	
Bond lengths (Å)	0.004 ± 0.0001
Bond angles (°)	0.60 ± 0.01
Impropers (°)	1.31 ± 0.04
Average pairwise coordinate RMSD (Å)	
Backbone	0.92 ± 0.30
Heavy atoms	1.41 ± 0.22

[a]Pairwise coordinate root-mean-square deviation (RMSD) was calculated for the 10 lowest-energy structures (regions 250–336 in U2AF2^{RRM2} and 31–43 in FUBP1$^{N\text{-}box}$) after water refinement. Ramachandran plot: 93.1%, 6.1%, 0.3%, and 0.4% of residues (regions 250–336 in U2AF2^{RRM2} and 31–43 in FUBP1$^{N\text{-}Box}$) are found in the most favored, additionally allowed, generously allowed, and disallowed regions.

Interestingly, both FUBP1 N-box-RRM interfaces show only limited interdigitation of the hydrophobic side chains, consistent with the modest binding affinity in the micromolar range.

In a recent survey of The Cancer Genome Atlas (TCGA), FUBP1 was noted for its particularly high rate of non-synonymous mutations in low-grade gliomas.[1] To learn about the mechanistic impact of such mutations, we systematically searched cancer mutation databases and identified 26 disease-related single-nucleotide variants (SNVs) within the FUBP1 N-box (Figure S4B). Five candidate mutations (A38D, A43E, K44R, I45F, and G47C) were selected by considering the magnitude of chemical shift changes occurring in the NMR titration of FUBP1$^{N\text{-}box}$ with U2AF2^{RRM2} (Figures S3B–S3D and S4B). In addition, we included L35V, which has been shown to weaken the FUBP1-PUF60 interaction.[61] NMR analysis revealed that A38D strongly impairs U2AF2 binding (Figures 3I and S4C–S4G). This is consistent with our structure in which A38 forms the core of the hydrophobic binding interface between FUBP1 N-box and U2AF2. A bulkier negatively charged side chain in this position is expected to introduce steric and electrostatic repulsion at the binding interface. Residue A38 in FUBP1 was also required for binding to PUF60 in a mutational study,[61] whereas L35V, which also affected the FUBP1-PUF60 interaction in that study, did not impair the interaction of FUBP1 with U2AF2 RRM2 (Figures S4C and S4D). A significant weakening of the U2AF2-FUBP1 interaction by A38D in the full-length context was also confirmed in cells using BRET (Figures 3C, 3J, and S2M). Here, some residual binding between FUBP1^{A38D} and U2AF2 was observed, probably because both proteins remain in proximity through binding to the same pre-mRNAs. As expected, A38D does not affect FUBP1-SF1 binding, which occurs via the A/B boxes (Figures 3C, S4H, and S4I). In summary, our experiments demonstrate that FUBP1 interacts directly with U2AF2 and SF1 via its N-terminal N-box and C-terminal A/B boxes, respectively. The former interaction is severely impaired by a cancer-associated mutation in FUBP1.

FUBP1 promotes U2AF2 binding to 3′ splice sites

To investigate the impact of FUBP1 on E complex formation, we monitored U2AF2 binding to RNA using *in vitro* iCLIP.[28] To this end, we designed a pool of short RNA transcripts (182 nt) representing ~2,000 natural 3′ splice sites from human transcripts, which we mixed with recombinant U2AF2^{RRM12} (see STAR Methods). Remarkably, addition of recombinant full-length FUBP1 (FUBP1^{FL}) results in stronger binding of U2AF2^{RRM12} to virtually all 3′ splice sites in the transcript pool (Figures 4A, 4B, and S5A–S5C; Table S1). The *in vivo* pattern of U2AF2 binding can thereby be reproduced *in vitro* in the presence of full-length FUBP1 (Figure 4C). The widespread effects are in contrast to those of our previous findings using *in vitro*-translated FUBP1, which affected only a few U2AF2 binding sites.[28] Hence, our updated experiments indicate that FUBP1 acts globally to stabilize U2AF2 binding. We find that this effect is dependent on FUBP1 concentration and is directly linked to the number of FUBP1 binding motifs upstream of the BP (Figure 4D). To confirm these findings in longer transcripts, we repeated the experiment with a pool of eight *in vitro* transcripts (2.0–5.7 kb; Figures S5D and S5E; Table S1). Indeed, addition of recombinant full-length FUBP1 increases the strength of U2AF2^{RRM12} binding at 3′ splice sites (Figures 4E and S5F) and thereby reproduces the *in vivo* binding pattern of U2AF2 (Figure 4F). Notably, this effect is considerably reduced with FUBP1$^{\Delta N}$ (impaired U2AF2 interaction), and it is completely abolished with FUBP1^{N74} (lacking KH domains). This highlights the importance of the N-box in FUBP1 for directly interacting with U2AF2 as well as of FUBP1's RNA binding for the stabilization of U2AF2 (Figures 4F and S5F). Together, this indicates that the interaction of FUBP1 with both pre-mRNA and U2AF2 globally promotes U2AF2 binding at the 3′ splice site during early spliceosomal assembly.

FUBP1 is critical for the splicing of long introns

To investigate the impact of FUBP1 on splicing, we generated a *FUBP1* knockout (KO) RPE1 cell line using CRISPR-Cas9 genome engineering (Figures 5A and S5G) and performed RNA-seq. *MYC* gene expression was unaltered, suggesting that it is not controlled by FUBP1 in RPE1 cells (Figure S5H). Next, we examined transcriptome-wide splicing and found 1,041 significant splicing changes, including 399 cassette exons (Figure 5B; Tables S1 and S3). Consistent with a role in splice site recognition, *FUBP1* KO preferentially leads to exon skipping (276 [69%] with delta percent spliced in [ΔPSI] < −0.1).

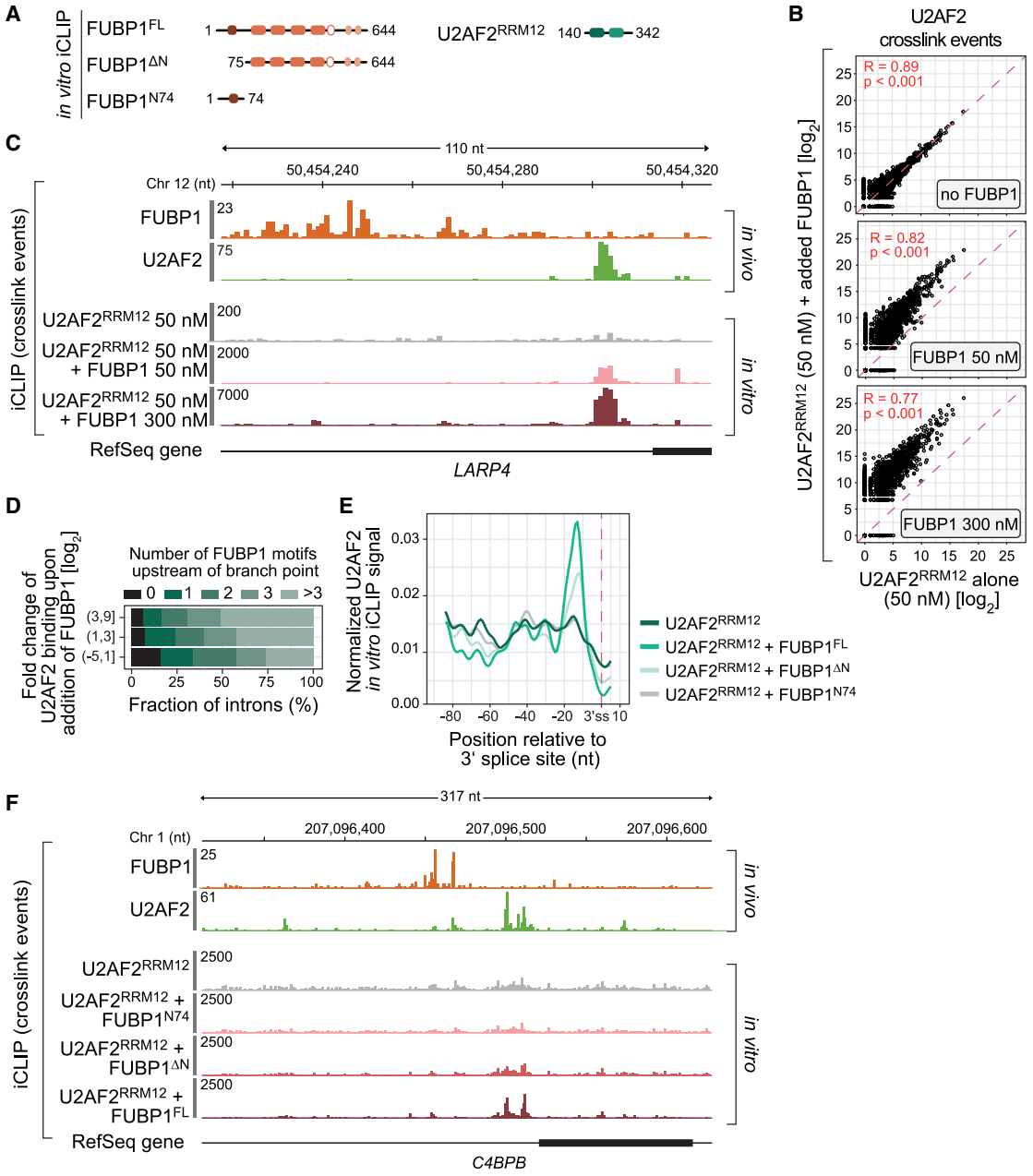


Figure 4. FUBP1 stabilizes U2AF2 binding at 3′ splice sites *in vitro*

(A) Overview of FUBP1 protein variants used in *in vitro* iCLIP experiments.

(B) Scatterplot of *in vitro* iCLIP signal in U2AF2 binding sites of U2AF2^{RRM12} alone and upon addition of full-length FUBP1 on a pool of 1,998 *in vitro* transcripts.

(C) Genome browser view of *LARP4* mRNA displaying *in vivo* iCLIP for FUBP1 and U2AF2 and *in vitro* iCLIP on the respective *in vitro* transcript for U2AF2 alone and after addition of full-length FUBP1.

(D) Number of FUBP1 binding motifs upstream of the BP ([−100 nt; −26 nt]) in relation to the log_2-transformed fold change of U2AF2^{RRM12} binding upon addition of full-length FUBP1 for 1,504 3′ splice sites in the *in vitro* transcripts.

(E) Metaprofile of U2AF2 binding at 3′ splice sites from *in vitro* iCLIP with long *in vitro* transcripts[28] and U2AF2^{RRM12} alone and after addition of FUBP1^{FL}, FUBP1^{N74}, or FUBP1$^{\Delta N}$. iCLIP signals were normalized by spike-in and averaged per nucleotide over all introns (n = 21).

(F) Genome browser view of *C4BPB* mRNA displaying *in vivo* iCLIP for FUBP1 and U2AF2 and *in vitro* iCLIP for U2AF2^{RRM12} alone and after addition of FUBP1^{FL}, FUBP1^{N74} or FUBP1$^{\Delta N}$.

CellPress
OPEN ACCESS

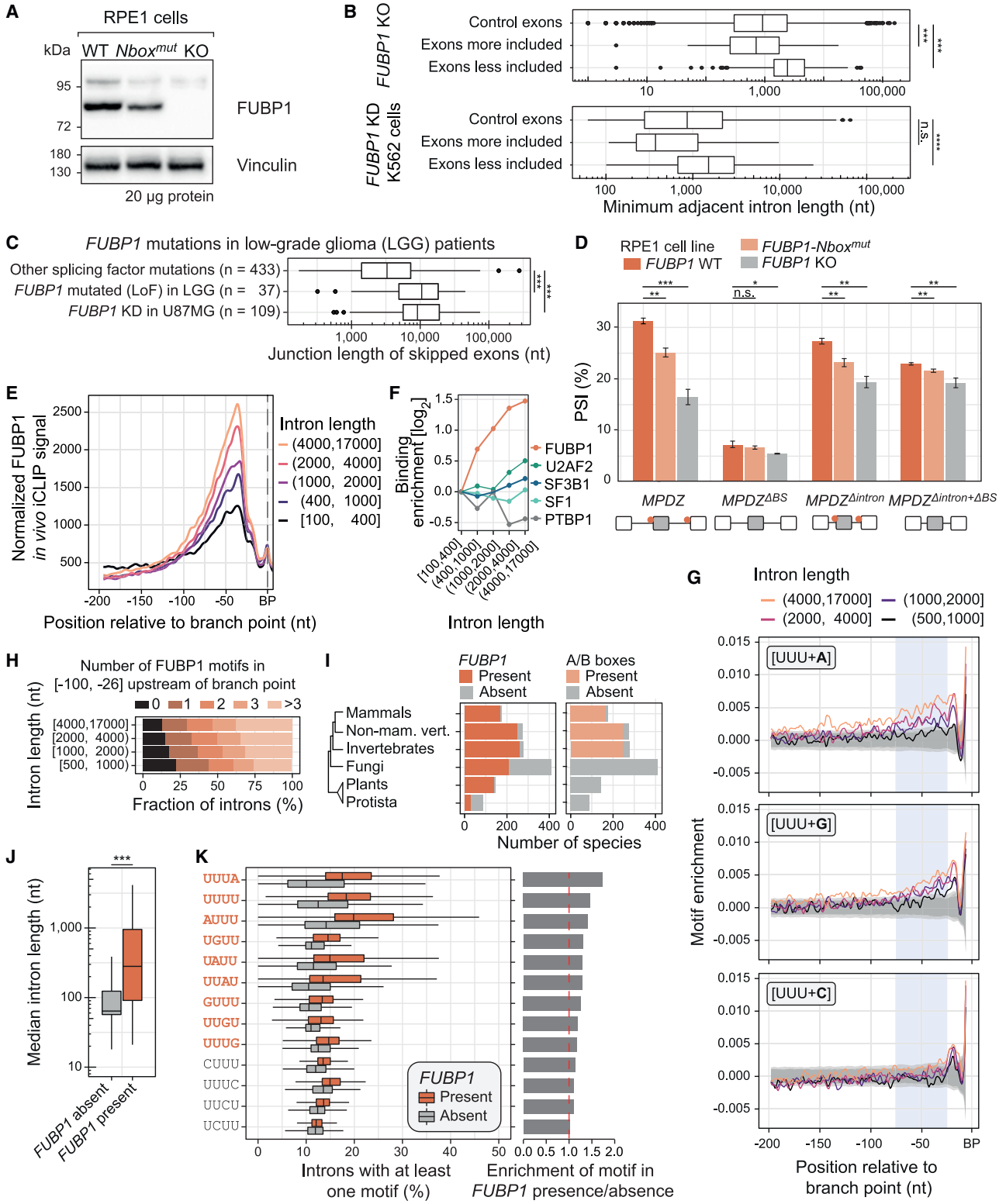


Figure 5. FUBP1 binds stronger to long introns and regulates exons flanked by long introns

(A) Western blot of FUBP1 in wild-type (WT), *FUBP1-Nbox*^{mut} mutant, and *FUBP1* KO RPE1 cells (Figure S5G). Vinculin acts as loading control.

(B) Minimum adjacent intron length for cassette exons more or less included upon *FUBP1* KO in RPE1 cells (n = 123/276) and *FUBP1* knockdown in K562 cells (n = 30/143) compared to unchanged control exons (RPE1, n = 10,301; K562, n = 1,910). ***p < 0.001, ****p < 0.0001, n.s., not significant.

(legend continued on next page)

A closer inspection revealed that the fate of an exon is related to the length of the flanking introns: decreased inclusion in *FUBP1* KO cells is typically observed for exons that are flanked by longer introns, compared with exons with increased or unchanged inclusion (Figure 5B, top). Most affected exons are alternative exons, but we observed the same effect for regulated constitutive exons (Figure S5I). Importantly, the effect on long introns can be recapitulated in ENCODE[62,63] data on *FUBP1* knockdown cells (Figure 5B, bottom). To test whether this depends on the interaction with U2AF2, we generated a *FUBP1-Nbox*^{mut} mutant with a targeted deletion of A38 and neighboring amino acids in the endogenous *FUBP1* gene in RPE1 cells (Figures 5A and S5G). Although overall fewer cassette exons are regulated in this mutant (n = 81), exons are predominantly skipped (n = 45), and these are flanked by longer introns (Figure S5J). Together, these data reveal that FUBP1 is important for the splicing of long introns and suggest a functional role for the N-box in this process.

To investigate whether *FUBP1* mutations in tumor cells affect splicing, we analyzed data from glioma patients.[1] Intriguingly, we found that skipped exons in patients with *FUBP1* loss-of-function mutations have longer adjacent introns than exons dysregulated in patients harboring other splicing factor mutations (Figures 5C and S6A). The effect is also evident upon *FUBP1* knockdown in the glioblastoma cell line U87MG from the same study (Figure 5C). Together, these data strongly suggest that FUBP1 plays a role in the efficient splicing of long introns, thereby affecting the inclusion of adjacent exons.

To validate the role of FUBP1 for long introns, we constructed a minigene for the alternative exon 18 in the *MPDZ* transcript, which is skipped upon *FUBP1* KO in RPE1 cells. The minigene comprises the alternative exon with the flanking constitutive exons and intervening long introns (>2.4 kb). *In vivo* iCLIP data show that FUBP1 binds at both 3′ splice sites, which was confirmed *in vitro* by EMSA with FUBP1$^{N\text{-box+KH}}$ (aa 1–457; Figures S6B and S6C). We observed a marked decrease of alternative exon inclusion from the *MPDZ* minigene in *FUBP1* KO (16% inclusion) and an intermediate effect (25%) in *FUBP1-Nbox*^{mut} cells, compared with wild-type (WT) cells (31% inclusion; Figures 5D, S6B, and S6D). Upon mutation of the FUBP1 binding sites, the exon showed reduced inclusion (7%) and did not change in the *FUBP1* KO. If the introns were shortened but the FUBP1 binding sites retained, the effect of *FUBP1* KO or mutation was reduced, albeit still present, consistent with the notion that the intron is still perceived as long due to the presence of FUBP1 binding site. By contrast, if the FUBP1 binding sites were also removed, exon inclusion no longer responded to *FUBP1* KO or *FUBP1-Nbox*^{mut}, highlighting that FUBP1 binding is specifically required for the long-intron variant.

Intriguingly, the changes at long introns are linked to FUBP1 binding. We found a substantial increase in FUBP1 binding at the 3′ splice sites of longer introns, both in absolute terms and relative to other splicing factors (Figures 5E and 5F). Differential FUBP1 binding was not observed for other exon-intron-related features, such as splice site, Py tract, and BP strength (Figures S6E–S6H). Furthermore, longer introns exhibit a marked enrichment of FUBP1 motifs upstream of the BP (Figures 5G and 5H). By contrast, random motif occurrences or splice site strength are independent of intron length (Figures S6I and S6J). Moreover, long introns were previously observed to preferentially locate to the nuclear periphery and exhibit a differential GC content architecture.[64,65] Indeed, we found that the occurrence of FUBP1 binding motifs correlates with the GC content architecture (Figures S6K–S6M). Furthermore, FUBP1 binds stronger to introns located in the nuclear periphery (Figure S6N) and to splice sites of exons with differential GC content architecture (Figures S7A–S7C). Further analysis indicated that both intron length and differential GC content architecture affect FUBP1 binding (Figure S7D).

Although splicing is an ancient molecular mechanism, gene architecture and especially intron length are subject to substantial evolutionary change (Figure S7E). We hypothesized that FUBP1 is present throughout Eukaryota and that lineage-specific losses or modifications of FUBP1 are accompanied by changes in average intron length. Indeed, we find overall that FUBP1 is well conserved. Although losses do occur, they are mostly observed in taxa with short introns such as protozoa and fungi (Figures 5I and 5J). Species with FUBP1 consistently harbor more FUBP1 motifs at their 3′ splice sites (Figure 5K). By contrast, U-rich motifs interspersed with C, which do not accumulate in the region of

(C) Junction length for less-included exons in RNA-seq from glioma patients with *FUBP1* loss-of-function (LoF) mutations, from a *FUBP1* siRNA knockdown in U87MG cells, and from *SF3B1/U2AF1/SRSF2* hotspot mutations and *RBM10* LoF mutation in different cancer patient samples. ***p < 0.001.

(D) Changes of exon inclusion (n = 3) in *FUBP1* WT, *FUBP1-Nbox*^{mut}, and *FUBP1* KO RPE1 cell lines upon intron shortening and/or removal of FUBP1 binding sites in the *MPDZ* minigene (Figure S6B). Data are represented as mean ± SD. Significance was determined by a two-sided Student's t-test with Benjamini-Hochberg correction. Red dots represent FUBP1 binding sites. *p < 0.05, **p < 0.01, ***p < 0.001, n.s., not significant.

(E) Metaprofile showing FUBP1 cross-link events relative to branch point for various intron lengths. iCLIP signals were normalized for expression and averaged per nucleotide over all introns.

(F) Quantification of binding signal based on area-under-the-curve (AUC) in main binding regions (see STAR Methods for details). Binding enrichment is defined as $\log_2$ fold change of AUC over AUC of introns with length in (100 nt, 400 nt).

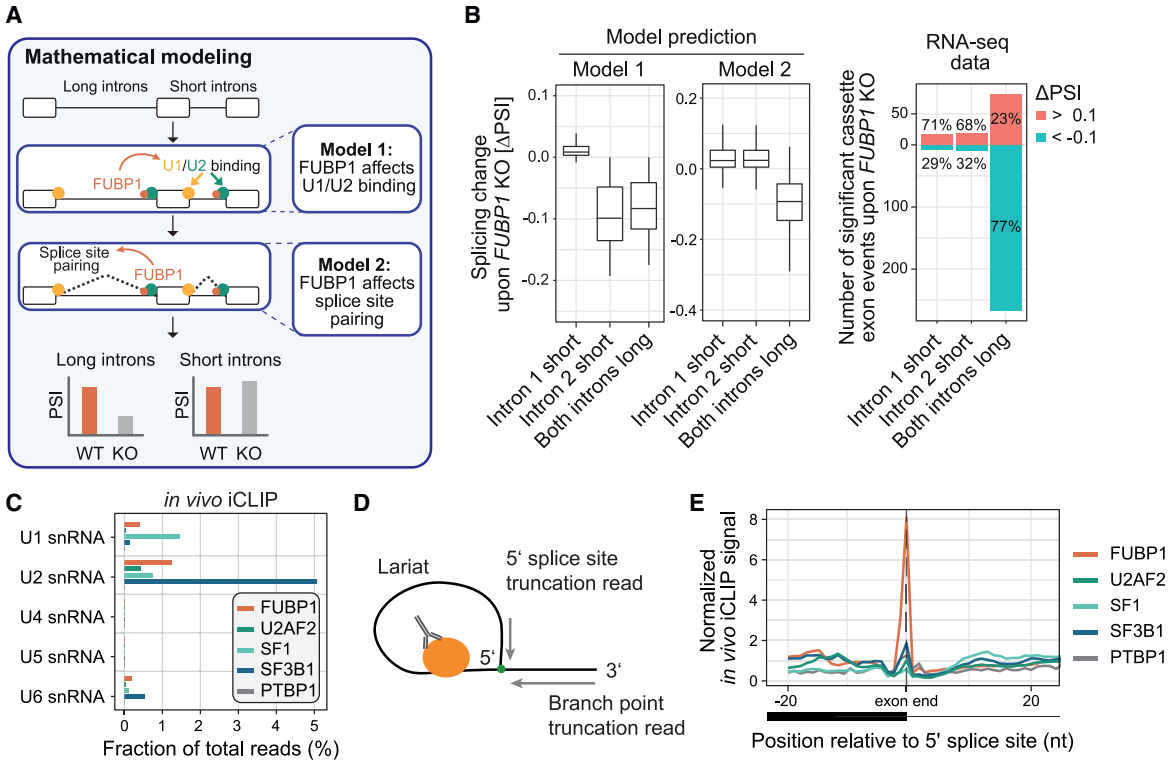
(G) Positional enrichment of FUBP1 binding motifs and control motifs relative to branch point and for various intron lengths. UUU+A/G/C, sets of four 4-mers containing UUU interspersed at any position with A/G/C. NNNN, 100,000 sets of random combinations of four 4-mers. 4-mer frequencies were calculated position-wise upstream of the BP and compared with average 4-mer frequencies in intronic control region.

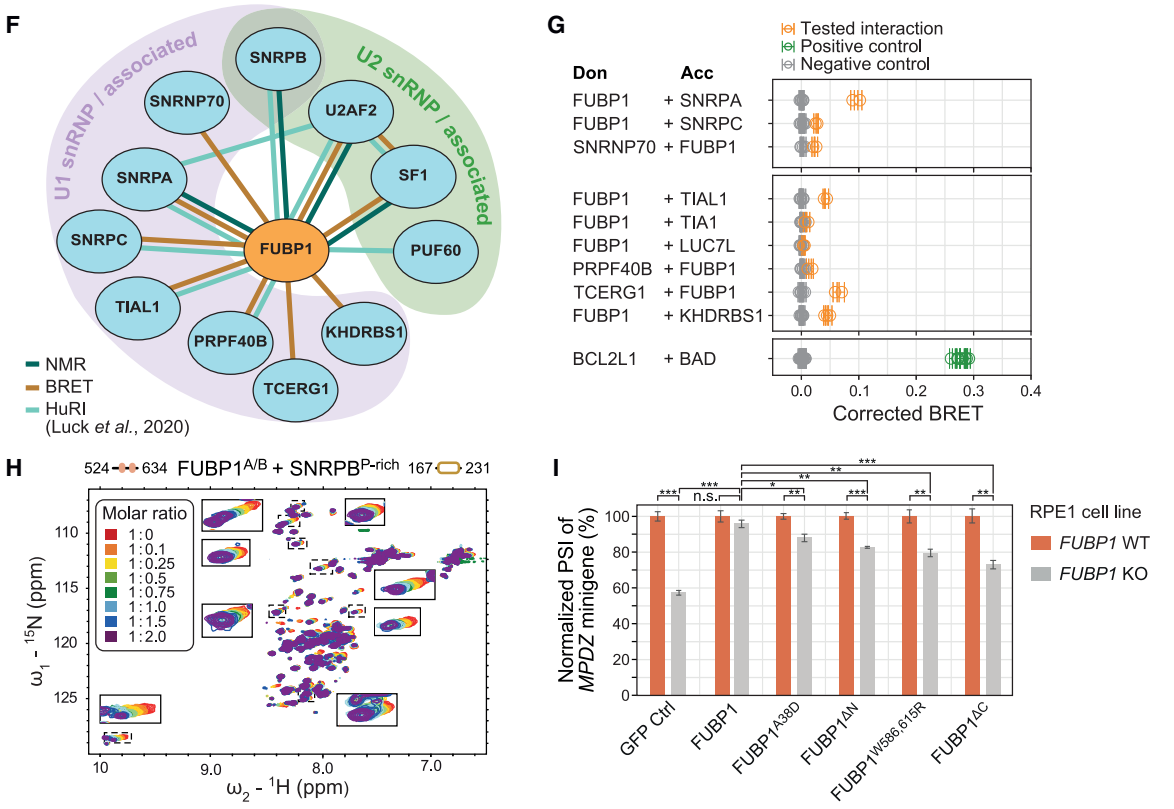
(H) Number of FUBP1 binding motifs upstream of the BP ([−100 nt; −26 nt]) for various intron lengths ([500, 1,000), n = 24,564 introns; [1,000, 2,000], n = 32,251 introns; [2,000, 4,000], n = 31,734 introns; [4,000, 17,000], n = 38,692).

(I) Phylogenetic profile of FUBP1. Tree indicates taxonomic range scanned for presence of FUBP1 orthologs. Fractions of species harboring ortholog to human FUBP1 (left) and carrying the A/B boxes (right) are shown.

(J) FUBP1 presence compared to median intron length per species. ***p < 0.001.

(K) Percentage of introns with at least one FUBP1 motif or control motifs present in 25-nt window located 25 nt upstream of the 3′ splice site.





(legend on next page)

FUBP1 binding (Figure 5G), are least enriched in species with FUBP1. Comparing FUBP1's domain architecture across eukaryotic evolution, we find that C-terminal A/B boxes are an animal-specific innovation. Their appearance in evolution is associated with an overall increase in intron length in animals compared with other eukaryotes (Figure 5I). Together, this suggests that FUBP1 binding to its RNA motifs and its protein-protein interfaces play important roles in the splicing of long introns.

FUBP1 interacts with both splice sites suggesting a function in cross-intron bridging

To decipher the molecular mechanism of FUBP1 action, we developed a kinetic model of cassette exon splicing using ordinary differential equations (Figures 6A and S7F; Table S4). In line with our previous work,[66] we considered a scenario for "exon definition" in which the U1 and U2 snRNPs recognize the 5′ and 3′ regions flanking an exon as functional units. The subsequent splice site pairing by U1/U2 snRNP interaction across the intron, that is, intron definition, triggers splicing catalysis, which results in either cassette exon inclusion, skipping, or intron retention in the model. We first simulated the loss of FUBP1 in a model in which FUBP1 solely acts on initial U1/U2 snRNP binding to exons (exon definition). However, our simulations argue against a pure exon definition effect, as the model cannot recapitulate the splicing changes that occur upon *FUBP1* KO (Figure 6B; model 1). According to our experimental data, exons flanked by two long introns are typically skipped upon *FUBP1* KO, whereas exons flanked by at least one short intron tend to show slightly increased inclusion. Surprisingly, the experimental data are more consistent with an alternative model in which FUBP1 enhances the pairing of splice sites across long (but not short) introns during intron definition. The model predicts reduced exon inclusion upon *FUBP1* KO specifically for exons flanked by two long introns, whereas exons flanked by one short intron moderately increase, irrespective of whether it is located upstream or downstream (Figure 6B; model 2). These results also hold true in a modified model, in which exons are not defined as functional units, and intron splicing solely requires U1 and U2 binding to flanking splice sites (Figure S7G, "intron definition model"). Taken together, the experimental observations are consistent with the kinetic model, which assumes that FUBP1 differentially affects long introns by promoting splice site pairing and the formation of catalytically active spliceosomes across long introns.

To test this prediction, we investigated the cross-linking of FUBP1 to snRNAs, indicative of its presence at different stages of splicing. First, FUBP1 showed substantial cross-linking to U2 snRNA, consistent with FUBP1 binding upstream of the BP where the U2 snRNP replaces SF1, indicating that FUBP1 is present during A-complex formation (Figure 6C). More importantly, FUBP1 also cross-links to U1 snRNA, which binds to the 5′ splice site, suggesting that FUBP1 is present during the bridging of the 3′ and 5′ splice sites, either during initial exon definition or also at later stages of intron definition. The latter is further supported by the cross-linking of FUBP1 to U6 snRNA, which replaces U1 snRNA at the 5′ splice site prior to lariat formation (Figure 6C). Hence, FUBP1 might be involved in intron bridging throughout the splicing cycle. We next searched our iCLIP datasets for evidence that FUBP1 is still bound in the spliceosomal C complex when the lariat has formed after the first splicing reaction. It has been shown that reads from the lariat truncate at the position where the 5′ splice site is covalently linked to the BP and is detected as a single-nucleotide-wide peak at the 5′ splice site (Figure 6D).[68,69] Indeed, we observed a strong peak in read truncations for FUBP1 at the 5′ splice site, whereas there was almost no signal for the other splicing factors tested (Figure 6E). This suggests that FUBP1 is present from the early stages of spliceosome assembly until at least the first catalytic step of the splicing reaction.

To further investigate whether FUBP1 is actively involved in splice-site bridging, we searched available binary protein-protein interaction data from yeast two-hybrid screens.[67] These data confirmed that FUBP1 binds to U2AF2 (Figure 6F). We also found evidence for FUBP1 interacting with several U1-associated proteins (SNRPA, SNRPC, TIAL1, and PRPF40B) as well

Figure 6. FUBP1 interacts with U1 snRNP components

(A) Kinetic model of FUBP1's effects on alternative splicing quantitatively describes steady-state abundance of splice products for a three-exon gene in control and *FUBP1* KO conditions. Two model variants were analyzed, in which FUBP1 affects the initial exon definition step near long introns (model 1), and the subsequent splicing reaction, promoting the excision of long introns (model 2). See STAR Methods for details.

(B) Simulated splicing changes upon *FUBP1* KO reflect transcriptome-wide RNA-seq data assuming that FUBP1 affects splicing catalysis (model 2). To reflect the heterogeneity of exons in the human transcriptome, kinetic parameters of the model were chosen at random, giving rise to an ensemble of 10,000 *in silico* exons. *FUBP1* KO was simulated for each *in silico* exon, assuming that FUBP1 either enhances exon definition (model 1) or the rate of splicing (model 2) for long (but not short) introns (see STAR Methods for details). In the data, significantly regulated cassette exons were classified based on flanking intron lengths (<400 nt = short, ≥400 nt = long).

(C) Fraction of total reads mapping to snRNAs using custom reference consisting of snRNAs (n = 10), tRNAs (n = 22), and rRNAs (n = 6).

(D) Schematic description of three-way junction of intron lariats. cDNAs can truncate not at the original protein-RNA interactions site but rather at the three-way junction. These cDNAs either start from the intron end and truncate at the BP or, alternatively, start downstream of the 5′ splice site and truncate at the first nucleotide of the intron.

(E) Metaprofiles showing cross-link events of FUBP1, U2AF2, SF3B1, SF1, and PTBP1 relative to the 5′ splice site. iCLIP signals were normalized for expression and averaged per nucleotide.

(F) Comprehensive interaction network of FUBP1 based on NMR, BRET, and published yeast two-hybrid data.[67]

(G) BRET measurements between FUBP1 and subunits of the U1 snRNP complex as well as U1 snRNP-associated proteins along with positive and negative control pairs. Biological replicates are shown. Error bars represent SD of technical triplicates.

(H) NMR titration of $FUBP1^{A/B}$ with $SNRPB^{P\text{-}rich}$ up to a molar ratio of 1:2. Significant chemical shift changes are highlighted by boxes.

(I) Percent-spliced-in (PSI) of *MPDZ* minigene upon transfection of WT and *FUBP1* KO RPE1 cells with different FUBP1 constructs. Data are represented as mean ± SD. Significance was determined by a two-sided Student's t-test with Benjamini-Hochberg correction. *p < 0.05, **p < 0.01, ***p < 0.001, n.s., not significant.

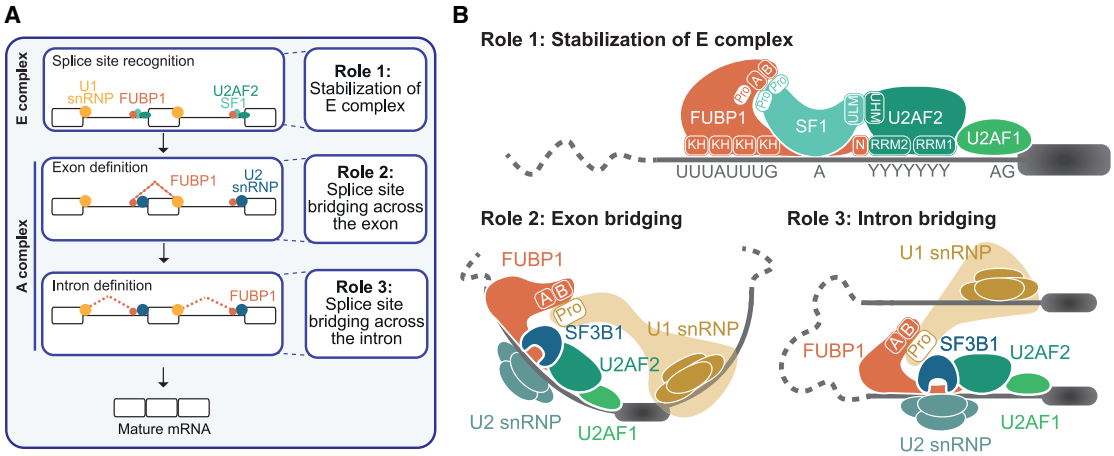


Figure 7. FUBP1 acts at multiple steps of early spliceosomal assembly

(A) The multiple roles of FUBP1 during spliceosomal complex assembly at the 3′ splice site.

(B) FUBP1 directly interacts with U2AF2, SF1, and additional U1/U2 snRNP components via distinct disordered interaction interfaces.

as with SNRPB, which is a member of the Sm protein ring in all snRNPs (Figure 6F). These and further interactions of FUBP1 with U1-associated proteins (TCERG1 and KHDRBS1) were confirmed using the BRET assay and/or NMR (Figures 6F, 6G, and S7H). Interestingly, several of the U1 snRNP-associated proteins harbor proline-rich regions, which potentially interact with the A/B boxes in FUBP1, similar to the FUBP1-SF1 interaction discussed above. Indeed, we observed significant changes in the NMR spectrum of FUBP1$^{A/B}$ upon the addition of a proline-rich peptide from SNRPB (Figure 6H), which were less pronounced with SNRPA and PRPF40B derivates (Figures S7I and S7J). This correlates well with the proline-rich region in SNRPB being much larger than in SNRPA or PRPF40B and thus avidity effects perhaps enhance the binding.

Finally, to confirm the importance of the FUBP1 A/B boxes and their role in splice-site bridging, we performed a complementation assay by expressing full-length GFP-FUBP1 and different mutants in both WT and *FUBP1* KO RPE1 cells. Effects on splicing were monitored using the co-transfected *MPDZ* minigene. As expected, GFP-FUBP1 complements the *FUBP1* KO cells and rescues *MPDZ* exon inclusion close to WT levels (Figures 6I, S7K, and S7L). Importantly, expression of GFP-FUBP1^{W586,615R} (mutations in the A/B boxes) or FUBP1$^{\Delta C}$ (complete deletion of the C terminus) impairs complementation in *FUBP1* KO cells. The same was also observed if the interaction with U2AF2 is perturbed by expressing either FUBP1^{A38D} (N-box mutation) or FUBP1$^{\Delta N}$ (complete deletion of the N terminus). Overall, these data demonstrate that both the A/B boxes and the N-box in FUBP1, which mediate the interactions with factors at the 5′ and 3′ splice sites, respectively, are functionally relevant for splicing.

DISCUSSION

FUBP1 is a general component of 3′ splice site definition

The recognition and pairing of splice sites, especially for the many long introns in the human transcriptome, are not well understood. In this study, we identified FUBP1 as a key component in 3′ splice site definition. We found that FUBP1 recognizes clustered U-rich elements interspersed by A or G that are present at virtually all 3′ splice sites and are most abundant for longer introns. Until now, four conserved intron-defining sequence motifs were known: the 5′ splice site motif, the BP sequence, the Py tract, and the 3′ splice site motif.[6] We propose the FUBP1 binding motif as a sequence signature that is relevant for spliceosomal assembly at long introns, which represent >80% of all human introns. Consistent with such a general role in splicing, FUBP1 has been detected in purified spliceosomes using mass spectrometry.[70–72]

We show that the four KH domains of FUBP1 recognize clustered arrays of binding motifs upstream of the BP. Multivalent interactions enhance binding affinity by avidity and enable the recognition of *cis*-elements in RNAs of variable length by combining individual KH-RNA motif interactions where multiple clustered RNA motifs may be separated by variable nucleotide linkers.[54] We find that the four KH domains are connected by flexible linkers, which facilitates scanning of extended RNA regions. The recognition of clustered RNA motifs by multidomain RBPs has been observed in IMP proteins and also involves four KH domains.[55] This suggests that KH domains working in concert might be a common mechanism for specifically recognizing clustered RNA motifs in extended RNA regions.

FUBP1 engages in multivalent interactions with 3′ and 5′ splice site components

We characterized two interfaces in FUBP1 that mediate protein-protein interactions: the N-box and the A/B boxes that are embedded in the intrinsically disordered N- and C-terminal regions of FUBP1, respectively. The N-box has been shown to interact with the RRM domain of PUF60 for regulation of transcription.[33,73,74] Here, we found that the FUBP1 N-box also binds to the RRM2 domain of U2AF2 and thereby mediates a functional interaction during pre-mRNA splicing. The N-box binds RRM2 opposite its RNA-binding surface, and thus, RNA

binding and FUBP1 binding do not compete. Notably, we have previously shown that the U2AF2 tandem domains adopt closed conformations and that RNA binding selects open arrangements.[15,29,75] Thus, binding of FUBP1 to the helical face of U2AF2 RRM2 might enhance RNA binding not only by stabilizing U2AF2 on the RNA but also by shifting the tandem RRM arrangements of U2AF2.

The A/B boxes of FUBP1 interact with intrinsically disordered proline-rich sequences within several U1 and U2 snRNP-associated proteins. This matches observations on the A/B boxes of the FUBP1 ortholog PSI in *Drosophila melanogaster*, which have been shown to bind to a proline-rich region in snRNP-U1-70K.[59] However, this region is not conserved in the human ortholog SNRNP70, and our BRET studies detected no such interaction between FUBP1 and SNRNP70. In general, linear motifs in proline-rich regions are recognized by structured regions such as WW or SH3 domains.[76] These interactions are generally weak but often enhanced by multivalent interactions.[77–81] Interestingly, the A/B boxes are unique to the FUBP family and appear to be unstructured regions in the ortholog PSI.[59] It will be interesting to learn how prevalent such an atypical mode of proline-rich sequence binding is and how it impacts cellular function.

FUBP1 contributes to spliceosome formation and guides the splicing of long introns

One important question is why FUBP1 is particularly relevant for long introns. Clearly, the splicing of long introns is difficult to achieve. For instance, it has been reported that exons flanking long introns are less included,[82,83] and that the splice sites of longer introns are stronger.[84,85] Consequently, longer introns require more complex regulation, such as the switch from initial exon definition to cross-intron spliceosomal complexes.[84,86] During exon definition, splice sites are recognized and paired across the exon, which is thereby defined as a functional unit. During the subsequent switch to intron definition, the complex shifts to a cross-intron pairing of splice sites (Figure 7). Our data suggest that FUBP1 acts at both steps. We propose that during exon definition, FUBP1 stabilizes U2AF2 and SF1 at the 3′ splice site. FUBP1 can thus strengthen the initial recognition of 3′ splice sites via its multivalent interactions with U2AF2, SF1, and pre-mRNA. The stabilization by FUBP1 and its interactions with the U1 snRNP across the exon might thus contribute to splice site recognition during exon definition.[86,87]

The interactions between FUBP1 and U1 snRNP components might also be relevant after the switch from exon definition to cross-intron pairing. Consistent with this model, we found that FUBP1 is still present at splice sites until the lariat is formed. In fact, FUBP1 forms cross-links to the U6 snRNA, which replaces U1 snRNA at the 5′ splice site. This indicates a role for FUBP1 in intron bridging during spliceosomal B-complex formation, particularly for long introns, as our experimental data and kinetic modeling suggest.

Several mechanisms and contributions to splice site bridging have been suggested, for example, the interactions between U1 and U2 snRNP proteins and RNA components[88–90] and the U2AF-associated RNA helicase UAP56.[91] It is conceivable that multiple contact sites act in concert to generate sufficient avidity to bring the splice sites together. Our data suggest that FUBP1—through multivalent interactions with pre-mRNA, proteins, and snRNAs located at the 5′ and 3′ splice sites—adds to these contacts throughout the splicing cycle. This is most pertinent for long introns harboring multiple FUBP1 *cis*-regulatory motifs.

In conclusion, we identify FUBP1 as a general splicing factor that ubiquitously binds at 3′ splice sites by means of a hitherto unknown *cis*-regulatory RNA sequence motif. The binding of FUBP1 and its interactions with multiple U1 and U2 snRNP components are pertinent to the efficient splicing of long introns.

Limitations of the study

Uridines are particularly prone to UV cross-linking, which can introduce bias to motif identification by iCLIP. However, we observed similar motifs using methods that do not involve UV cross-linking (NMR spectroscopy, ITC, and EMSA); therefore, we are confident that our conclusions in this regard are valid.

Upon depletion of FUBP1 in our KO or knockdown cell lines, other factors (such as the close paralog KHSRP) might, to some extent, take on the role of FUBP1. Together with cellular quality control mechanisms that degrade mis-spliced transcripts, this might reduce the effects of FUBP1 perturbation that we observed in our RNA-seq analysis. We might clarify such effects in the future by combing acute depletion of FUBP1 by means of degron tags with analysis of nascent RNA.

U2AF2 RRM2 and FUBP1 N-box interact with weak affinity in the micromolar range. Although it is likely that the simultaneous binding of U2AF2 and FUBP1 to the RNA further stabilizes this interaction, we cannot exclude the involvement of other factors.

In general, introns may be characterized by a multitude of features, among which length is just one. For example, intron length is known to correlate with elevated differential GC content and overall lower intron and exon GC content.[65] In addition, genes with longer introns have been shown to preferentially localize to the nuclear periphery,[64] and their transcripts therefore might interact with different splicing factors than for genes at the nuclear center. The question of whether these attributes rather complement each other or are causally related remains to be answered.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - RPE1 cell lines and culture conditions
  - HeLa cell line and culture conditions
  - HEK cell line and culture conditions
  - Recombinant protein expression
- METHOD DETAILS
  - Establishing *FUBP1* KO/*Nbox*^{mut} cell lines

- Immunoblotting
- RPE1 RNA-seq
- HeLa RNA-seq
- Semi-quantitative RT-PCR
- *In vivo* iCLIP
- *In vitro* iCLIP
- Protein expression and purification
- NMR spectroscopy
- *In vitro* binding assays
- BRET

- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Preprocessing of RNA-seq data
  - Preprocessing of *in vivo* iCLIP data
  - Metaprofiles for *in vivo* iCLIP data
  - iCLIP binding site definition (peak calling)
  - Saturation analysis
  - Motif enrichment for *in vivo* iCLIP
  - Motif enrichment upstream of branch points
  - Abundance of FUBP1 motif at 3′ splice sites
  - Analysis of *in vitro* iCLIP data
  - Intron length analyses of RNA-seq data
  - ENCODE data analysis
  - Splicing changes upon FUBP1 LoF mutations
  - Mutations in FUBP1 in cancer patients
  - Scoring of splice site features
  - Evolutionary analyses
  - Analysis of RBP crosslinking to snRNAs
  - Subnuclear distribution of FUBP1-bound genes
  - Mathematical modeling

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.molcel.2023.07.002.

ACKNOWLEDGMENTS

We thank all the members of the Luck, Sattler, and König labs for their help and discussion. We thank Malgorzata Rogalska and Juan Valcárcel for discussions and comments on the manuscript, Philipp Trepte and the Wanker group for sharing protocols and reagents and for help in setting up BRET assays, Christian Schäfer for help with BRET assays, Eric Schumbera for help with BRET data processing, Fridolin Kielisch for help with statistical analyses, Mario Keller for bioinformatics advice, André Mourão for SNRPB$^{P\text{-}rich}$ plasmid, Sam Asami and Gerd Gemmecker for support with NMR experiments, Manuel Kaulich for reagents, and Chris Smith and Jernej Ule for PTBP1-RB40 antibody and resequencing. We thank Adrian Neal for editing and commenting on the manuscript. We thank the Core Facilities at IMB, in particular Protein Production, Microscopy, Bioinformatics, Genomics, and Flow Cytometry.

We acknowledge IMB Genomics Core Facility and its NextSeq 500 sequencer (funded by the Deutsche Forschungsgemeinschaft [DFG, German Research Foundation] INST 247/870-1 FUGG) and access to NMR spectrometers at Bavarian NMR Center. This work was supported by DFG grants to K.L. (LU 2568/1-1; SFB1551 Project no. 464588647), J.K. (SPP1935 Project no. 273941853, KO4566/2-1, SFB1551 Project No. 464588647, TRR 319 Project no. 439669440, and GRK2526/1 Project no. 407023052), K.Z. (SPP1935 Project no. 273941853), S.L. (LE 3473/2–3), and M.S. (SPP1935 Project no. 273941853, SA823/10-1, and SFB1035 Project no. 201302640). C.H. acknowledges the Fonds der Chemischen Industrie for Kekulé fellowship, and S.M.-L. acknowledges EU Horizon 2020 Research and Innovation program under the Marie Skłodovska-Curie grant agreement No. 792692. J.S. acknowledges a PhD stipend from IMB's collaborative research initiative.

AUTHOR CONTRIBUTIONS

S.E., C.B., A. Busch, and A.D.L. performed the bioinformatic analyses. C.H., H.-S.K., and S.M.-L. performed the structural, biophysical, and biochemical experiments and analyses. M.M. Mulorz, A. Buchbender, F.X.R.S., L.L.A., H.H., K.T., and M.M. Möckel performed the functional genomics, *in vitro* iCLIP, and minigene reporter experiments. D.H., J.S., and M.W. performed the BRET experiments. P.K. and S.L. performed the mathematical modeling. I.E. performed the evolutionary analysis. S.E., C.H., M.M. Mulorz, K.Z., K.L., M.S., and J.K. designed the study and wrote the manuscript. All authors read and commented on the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: January 4, 2023
Revised: May 19, 2023
Accepted: July 3, 2023
Published: July 27, 2023

REFERENCES

1. Seiler, M., Peng, S., Agrawal, A.A., Palacino, J., Teng, T., Zhu, P., Smith, P.G., Cancer; Genome; Atlas; Research Network, Buonamici, S., and Yu, L. (2018). Somatic mutational landscape of splicing factor genes and their functional consequences across 33 cancer types. Cell Rep. *23*, 282–296.e4. https://doi.org/10.1016/j.celrep.2018.01.088.
2. Bonnal, S.C., López-Oreja, I., and Valcárcel, J. (2020). Roles and mechanisms of alternative splicing in cancer – implications for care. Nat. Rev. Clin. Oncol. *17*, 457–474. https://doi.org/10.1038/s41571-020-0350-x.
3. Gebauer, F., Schwarzl, T., Valcárcel, J., and Hentze, M.W. (2021). RNA-binding proteins in human genetic disease. Nat. Rev. Genet. *22*, 185–198. https://doi.org/10.1038/s41576-020-00302-y.
4. Shi, Y. (2017). Mechanistic insights into precursor messenger RNA splicing by the spliceosome. Nat. Rev. Mol. Cell Biol. *18*, 655–670. https://doi.org/10.1038/nrm.2017.86.
5. Wilkinson, M.E., Charenton, C., and Nagai, K. (2020). RNA splicing by the spliceosome. Annu. Rev. Biochem. *89*, 359–388. https://doi.org/10.1146/annurev-biochem-091719-064225.
6. Wahl, M.C., Will, C.L., and Lührmann, R. (2009). The spliceosome: design principles of a dynamic RNP machine. Cell *136*, 701–718. https://doi.org/10.1016/j.cell.2009.02.009.
7. Papasaikas, P., and Valcárcel, J. (2016). The spliceosome: the ultimate RNA chaperone and sculptor. Trends Biochem. Sci. *41*, 33–45. https://doi.org/10.1016/j.tibs.2015.11.003.
8. Berglund, J.A., Abovich, N., and Rosbash, M. (1998). A cooperative interaction between U2AF65 and mBBP/SF1 facilitates branchpoint region recognition. Genes Dev. *12*, 858–867. https://doi.org/10.1101/gad.12.6.858.
9. Liu, Z., Luyten, I., Bottomley, M.J., Messias, A.C., Houngninou-Molango, S., Sprangers, R., Zanier, K., Krämer, A., and Sattler, M. (2001). Structural basis for recognition of the intron branch site RNA by splicing factor 1. Science *294*, 1098–1102. https://doi.org/10.1126/science.1064719.
10. Selenko, P., Gregorovic, G., Sprangers, R., Stier, G., Rhani, Z., Krämer, A., and Sattler, M. (2003). Structural basis for the molecular recognition between human splicing factors U2AF65 and SF1/mBBP. Mol. Cell *11*, 965–976. https://doi.org/10.1016/s1097-2765(03)00115-1.
11. Kielkopf, C.L., Rodionova, N.A., Green, M.R., and Burley, S.K. (2001). A novel peptide recognition mode revealed by the X-ray structure of a core U2AF35/U2AF65 heterodimer. Cell *106*, 595–605. https://doi.org/10.1016/s0092-8674(01)00480-9.
12. Wu, S., Romfo, C.M., Nilsen, T.W., and Green, M.R. (1999). Functional recognition of the 3′ splice site AG by the splicing factor U2AF35. Nature *402*, 832–835. https://doi.org/10.1038/45590.

13. Merendino, L., Guth, S., Bilbao, D., Martínez, C., and Valcárcel, J. (1999). Inhibition of msl-2 splicing by Sex-lethal reveals interaction between U2AF35 and the 3′ splice site AG. Nature *402*, 838–841. https://doi.org/10.1038/45602.

14. Agrawal, A.A., Salsi, E., Chatrikhi, R., Henderson, S., Jenkins, J.L., Green, M.R., Ermolenko, D.N., and Kielkopf, C.L. (2016). An extended U2AF(65)–RNA-binding domain recognizes the 3′ splice site signal. Nat. Commun. *7*, 10950. https://doi.org/10.1038/ncomms10950.

15. Mackereth, C.D., Madl, T., Bonnal, S., Simon, B., Zanier, K., Gasch, A., Rybin, V., Valcárcel, J., and Sattler, M. (2011). Multi-domain conformational selection underlies pre-mRNA splicing regulation by U2AF. Nature *475*, 408–411. https://doi.org/10.1038/nature10171.

16. Zamore, P.D., and Green, M.R. (1989). Identification, purification, and biochemical characterization of U2 small nuclear ribonucleoprotein auxiliary factor. Proc. Natl. Acad. Sci. USA *86*, 9243–9247. https://doi.org/10.1073/pnas.86.23.9243.

17. Berglund, J.A., Chua, K., Abovich, N., Reed, R., and Rosbash, M. (1997). The splicing factor BBP interacts specifically with the pre-mRNA branchpoint sequence UACUAAC. Cell *89*, 781–787. https://doi.org/10.1016/s0092-8674(00)80261-5.

18. Crisci, A., Raleff, F., Bagdiul, I., Raabe, M., Urlaub, H., Rain, J.-C., and Krämer, A. (2015). Mammalian splicing factor SF1 interacts with SURP domains of U2 snRNP-associated proteins. Nucleic Acids Res. *43*, 10456–10473. https://doi.org/10.1093/nar/gkv952.

19. Wahl, M.C., and Lührmann, R. (2015). SnapShot: spliceosome dynamics I. Cell *161*, 1474–1474e1. https://doi.org/10.1016/j.cell.2015.05.050.

20. Tholen, J., and Galej, W.P. (2022). Structural studies of the spliceosome: bridging the gaps. Curr. Opin. Struct. Biol. *77*, 102461. https://doi.org/10.1016/j.sbi.2022.102461.

21. Ule, J., and Blencowe, B.J. (2019). Alternative splicing regulatory networks: functions, mechanisms, and evolution. Mol. Cell *76*, 329–345. https://doi.org/10.1016/j.molcel.2019.09.017.

22. Zuo, P., and Maniatis, T. (1996). The splicing factor U2AF35 mediates critical protein-protein interactions in constitutive and enhancer-dependent splicing. Genes Dev. *10*, 1356–1368. https://doi.org/10.1101/gad.10.11.1356.

23. Saulière, J., Sureau, A., Expert-Bezançon, A., and Marie, J. (2006). The polypyrimidine tract binding protein (PTB) represses splicing of exon 6B from the beta-tropomyosin pre-mRNA by directly interfering with the binding of the U2AF65 subunit. Mol. Cell. Biol. *26*, 8755–8769. https://doi.org/10.1128/MCB.00893-06.

24. Soares, L.M.M., Zanier, K., Mackereth, C., Sattler, M., and Valcárcel, J. (2006). Intron removal requires proofreading of U2AF/3′ splice site recognition by DEK. Science *312*, 1961–1965. https://doi.org/10.1126/science.1128659.

25. Warf, M.B., Diegel, J.V., von Hippel, P.H., and Berglund, J.A. (2009). The protein factors MBNL1 and U2AF65 bind alternative RNA structures to regulate splicing. Proc. Natl. Acad. Sci. USA *106*, 9203–9208. https://doi.org/10.1073/pnas.0900342106.

26. Tavanez, J.P., Madl, T., Kooshapur, H., Sattler, M., and Valcárcel, J. (2012). hnRNP A1 proofreads 3′ splice site recognition by U2AF. Mol. Cell *45*, 314–329. https://doi.org/10.1016/j.molcel.2011.11.033.

27. Zarnack, K., König, J., Tajnik, M., Martincorena, I., Eustermann, S., Stévant, I., Reyes, A., Anders, S., Luscombe, N.M., and Ule, J. (2013). Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. Cell *152*, 453–466. https://doi.org/10.1016/j.cell.2012.12.023.

28. Sutandy, F.X.R., Ebersberger, S., Huang, L., Busch, A., Bach, M., Kang, H.-S., Fallmann, J., Maticzka, D., Backofen, R., Stadler, P.F., et al. (2018). In vitro iCLIP-based modeling uncovers how the splicing factor U2AF2 relies on regulation by cofactors. Genome Res. *28*, 699–713. https://doi.org/10.1101/gr.229757.117.

29. Voith von Voithenberg, L., Sánchez-Rico, C., Kang, H.-S., Madl, T., Zanier, K., Barth, A., Warner, L.R., Sattler, M., and Lamb, D.C. (2016). Recognition of the 3′ splice site RNA by the U2AF heterodimer involves a dynamic population shift. Proc. Natl. Acad. Sci. USA *113*. E7169–E7175. https://doi.org/10.1073/pnas.1605873113.

30. Kang, H.-S., Sánchez-Rico, C., Ebersberger, S., Sutandy, F.X.R., Busch, A., Welte, T., Stehle, R., Hipp, C., Schulz, L., Buchbender, A., et al. (2020). An autoinhibitory intramolecular interaction proof-reads RNA recognition by the essential splicing factor U2AF2. Proc. Natl. Acad. Sci. USA *117*, 7140–7149. https://doi.org/10.1073/pnas.1913483117.

31. Debaize, L., and Troadec, M.-B. (2019). The master regulator FUBP1: its emerging role in normal cell function and malignant development. Cell. Mol. Life Sci. *76*, 259–281. https://doi.org/10.1007/s00018-018-2933-6.

32. Duncan, R., Bazar, L., Michelotti, G., Tomonaga, T., Krutzsch, H., Avigan, M., and Levens, D. (1994). A sequence-specific, single-strand binding protein activates the far upstream element of c-myc and defines a new DNA-binding motif. Genes Dev. *8*, 465–480. https://doi.org/10.1101/gad.8.4.465.

33. Liu, J., Kouzine, F., Nie, Z., Chung, H.-J., Elisha-Feil, Z., Weber, A., Zhao, K., and Levens, D. (2006). The FUSE/FBP/FIR/TFIIH system is a molecular machine programming a pulse of *c-myc* expression. EMBO J. *25*, 2119–2130. https://doi.org/10.1038/sj.emboj.7601101.

34. Cukier, C.D., Hollingworth, D., Martin, S.R., Kelly, G., Díaz-Moreno, I., and Ramos, A. (2010). Molecular basis of FIR-mediated *c-myc* transcriptional control. Nat. Struct. Mol. Biol. *17*, 1058–1064. https://doi.org/10.1038/nsmb.1883.

35. Li, H., Wang, Z., Zhou, X., Cheng, Y., Xie, Z., Manley, J.L., and Feng, Y. (2013). Far upstream element-binding protein 1 and RNA secondary structure both mediate second-step splicing repression. Proc. Natl. Acad. Sci. USA *110*. E2687–E2695. https://doi.org/10.1073/pnas.1310607110.

36. Hwang, I., Cao, D., Na, Y., Kim, D.-Y., Zhang, T., Yao, J., Oh, H., Hu, J., Zheng, H., Yao, Y., and Paik, J. (2018). Far upstream element-binding protein 1 regulates LSD1 alternative splicing to promote terminal differentiation of neural progenitors. Stem Cell Reports *10*, 1208–1221. https://doi.org/10.1016/j.stemcr.2018.02.013.

37. Jacob, A.G., Singh, R.K., Mohammad, F., Bebee, T.W., and Chandler, D.S. (2014). The splicing factor FUBP1 is required for the efficient splicing of oncogene MDM2 pre-mRNA. J. Biol. Chem. *289*, 17350–17364. https://doi.org/10.1074/jbc.M114.554717.

38. Miro, J., Laaref, A.M., Rofidal, V., Lagrafeuille, R., Hem, S., Thorel, D., Méchin, D., Mamchaoui, K., Mouly, V., Claustres, M., and Tuffery-Giraud, S. (2015). FUBP1: a new protagonist in splicing regulation of the DMD gene. Nucleic Acids Res. *43*, 2378–2389. https://doi.org/10.1093/nar/gkv086.

39. Ni, X., Knapp, S., and Chaikuad, A. (2020). Comparative structural analyses and nucleotide-binding characterization of the four KH domains of FUBP1. Sci. Rep. *10*, 13459. https://doi.org/10.1038/s41598-020-69832-z.

40. Wang, H., Zhang, R., Li, E., Yan, R., Ma, B., and Ma, Q. (2022). Pan-cancer transcriptome and immune infiltration analyses reveal the oncogenic role of far upstream element-binding protein 1 (FUBP1). Front. Mol. Biosci. *9*, 794715. https://doi.org/10.3389/fmolb.2022.794715.

41. Elman, J.S., Ni, T.K., Mengwasser, K.E., Jin, D., Wronski, A., Elledge, S.J., and Kuperwasser, C. (2019). Identification of FUBP1 as a long tail cancer driver and widespread regulator of tumor suppressor and oncogene alternative splicing. Cell Rep. *28*, 3435–3449.e5. https://doi.org/10.1016/j.celrep.2019.08.060.

42. Wang, J., Schultz, P.G., and Johnson, K.A. (2018). Mechanistic studies of a small-molecule modulator of SMN2 splicing. Proc. Natl. Acad. Sci. USA *115*. E4604–E4612. https://doi.org/10.1073/pnas.1800260115.

43. König, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D.J., Luscombe, N.M., and Ule, J. (2010). iCLIP reveals the function of

hnRNP particles in splicing at individual nucleotide resolution. Nat. Struct. Mol. Biol. *17*, 909–915. https://doi.org/10.1038/nsmb.1838.

44. Buchbender, A., Mutter, H., Sutandy, F.X.R., Körtel, N., Hänel, H., Busch, A., Ebersberger, S., and König, J. (2020). Improved library preparation with the new iCLIP2 protocol. Methods *178*, 33–48. https://doi.org/10.1016/j.ymeth.2019.10.003.

45. Valcárcel, J., Gaur, R.K., Singh, R., and Green, M.R. (1996). Interaction of U2AF65 RS region with pre-mRNA branch point and promotion of base pairing with U2 snRNA [corrected]. Science *273*, 1706–1709. https://doi.org/10.1126/science.273.5282.1706.

46. Singh, R., Valcárcel, J., and Green, M.R. (1995). Distinct binding specificities and functions of higher eukaryotic polypyrimidine tract-binding proteins. Science *268*, 1173–1176. https://doi.org/10.1126/science.7761834.

47. Sugimoto, Y., König, J., Hussain, S., Zupan, B., Curk, T., Frye, M., and Ule, J. (2012). Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. Genome Biol. *13*, R67. https://doi.org/10.1186/gb-2012-13-8-r67.

48. Gozani, O., Potashkin, J., and Reed, R. (1998). A potential role for U2AF-SAP 155 interactions in recruiting U2 snRNP to the branch site. Mol. Cell. Biol. *18*, 4752–4760. https://doi.org/10.1128/MCB.18.8.4752.

49. Xue, Y., Zhou, Y., Wu, T., Zhu, T., Ji, X., Kwon, Y.-S., Zhang, C., Yeo, G., Black, D.L., Sun, H., et al. (2009). Genome-wide analysis of PTB-RNA interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping. Mol. Cell *36*, 996–1006. https://doi.org/10.1016/j.molcel.2009.12.003.

50. Llorian, M., Schwartz, S., Clark, T.A., Hollander, D., Tan, L.-Y., Spellman, R., Gordon, A., Schweitzer, A.C., de la Grange, P., Ast, G., and Smith, C.W.J. (2010). Position-dependent alternative splicing activity revealed by global profiling of alternative splicing events regulated by PTB. Nat. Struct. Mol. Biol. *17*, 1114–1123. https://doi.org/10.1038/nsmb.1881.

51. Shao, C., Yang, B., Wu, T., Huang, J., Tang, P., Zhou, Y., Zhou, J., Qiu, J., Jiang, L., Li, H., et al. (2014). Mechanisms for U2AF to define 3′ splice sites and regulate alternative splicing in the human genome. Nat. Struct. Mol. Biol. *21*, 997–1005. https://doi.org/10.1038/nsmb.2906.

52. Valverde, R., Edwards, L., and Regan, L. (2008). Structure and function of KH domains. FEBS J. *275*, 2712–2726. https://doi.org/10.1111/j.1742-4658.2008.06411.x.

53. Fukumura, K., Yoshimoto, R., Sperotto, L., Kang, H.-S., Hirose, T., Inoue, K., Sattler, M., and Mayeda, A. (2021). SPF45/RBM17-dependent, but not U2AF-dependent, splicing in a distinct subset of human short introns. Nat. Commun. *12*, 4910. https://doi.org/10.1038/s41467-021-24879-y.

54. Mackereth, C.D., and Sattler, M. (2012). Dynamics in multi-domain protein recognition of RNA. Curr. Opin. Struct. Biol. *22*, 287–296. https://doi.org/10.1016/j.sbi.2012.03.013.

55. Schneider, T., Hung, L.-H., Aziz, M., Wilmen, A., Thaum, S., Wagner, J., Janowski, R., Müller, S., Schreiner, S., Friedhoff, P., et al. (2019). Combinatorial recognition of clustered RNA elements by the multidomain RNA-binding protein IMP3. Nat. Commun. *10*, 2266. https://doi.org/10.1038/s41467-019-09769-8.

56. Siomi, H., Matunis, M.J., Michael, W.M., and Dreyfuss, G. (1993). The pre-mRNA binding K protein contains a novel evolutionarily conserved motif. Nucleic Acids Res. *21*, 1193–1198. https://doi.org/10.1093/nar/21.5.1193.

57. Beuth, B., García-Mayoral, M.F., Taylor, I.A., and Ramos, A. (2007). Scaffold-independent analysis of RNA-protein interactions: the Nova-1 KH3-RNA complex. J. Am. Chem. Soc. *129*, 10205–10210. https://doi.org/10.1021/ja072365q.

58. Trepte, P., Kruse, S., Kostova, S., Hoffmann, S., Buntru, A., Tempelmeier, A., Secker, C., Diez, L., Schulz, A., Klockmeier, K., et al. (2018). LuTHy: a double-readout bioluminescence-based two-hybrid technology for quantitative mapping of protein-protein interactions in mammalian cells. Mol. Syst. Biol. *14*, e8071. https://doi.org/10.15252/msb.20178071.

59. Ignjatovic, T., Yang, J.-C., Butler, J., Neuhaus, D., and Nagai, K. (2005). Structural basis of the interaction between P-element somatic inhibitor and U1-70k essential for the alternative splicing of P-element transposase. J. Mol. Biol. *351*, 52–65. https://doi.org/10.1016/j.jmb.2005.04.077.

60. Labourier, E., Adams, M.D., and Rio, D.C. (2001). Modulation of P-element pre-mRNA splicing by a direct interaction between PSI and U1 snRNP 70K protein. Mol. Cell *8*, 363–373. https://doi.org/10.1016/s1097-2765(01)00311-2.

61. Chung, H.-J., Liu, J., Dundr, M., Nie, Z., Sanford, S., and Levens, D. (2006). FBPs are calibrated molecular tools to adjust gene expression. Mol. Cell. Biol. *26*, 6584–6597. https://doi.org/10.1128/MCB.00754-06.

62. ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74. https://doi.org/10.1038/nature11247.

63. Luo, Y., Hitz, B.C., Gabdank, I., Hilton, J.A., Kagda, M.S., Lam, B., Myers, Z., Sud, P., Jou, J., Lin, K., et al. (2020). New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. Nucleic Acids Res. *48*. D882–D889. https://doi.org/10.1093/nar/gkz1062.

64. Tammer, L., Hameiri, O., Keydar, I., Roy, V.R., Ashkenazy-Titelman, A., Custódio, N., Sason, I., Shayevitch, R., Rodríguez-Vaello, V., Rino, J., et al. (2022). Gene architecture directs splicing outcome in separate nuclear spatial regions. Mol. Cell *82*, 1021–1034.e8. https://doi.org/10.1016/j.molcel.2022.02.001.

65. Amit, M., Donyo, M., Hollander, D., Goren, A., Kim, E., Gelfman, S., Lev-Maor, G., Burstein, D., Schwartz, S., Postolsky, B., et al. (2012). Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. Cell Rep. *1*, 543–556. https://doi.org/10.1016/j.celrep.2012.03.013.

66. Enculescu, M., Braun, S., Thonta Setty, S., Busch, A., Zarnack, K., König, J., and Legewie, S. (2020). Exon definition facilitates reliable control of alternative splicing in the RON proto-oncogene. Biophys. J. *118*, 2027–2041. https://doi.org/10.1016/j.bpj.2020.02.022.

67. Luck, K., Kim, D.-K., Lambourne, L., Spirohn, K., Begg, B.E., Bian, W., Brignall, R., Cafarelli, T., Campos-Laborie, F.J., Charloteaux, B., et al. (2020). A reference map of the human binary protein interactome. Nature *580*, 402–408. https://doi.org/10.1038/s41586-020-2188-x.

68. Briese, M., Haberman, N., Sibley, C.R., Faraway, R., Elser, A.S., Chakrabarti, A.M., Wang, Z., König, J., Perera, D., Wickramasinghe, V.O., et al. (2019). A systems view of spliceosomal assembly and branchpoints with iCLIP. Nat. Struct. Mol. Biol. *26*, 930–940. https://doi.org/10.1038/s41594-019-0300-4.

69. Cordiner, R.A., Dou, Y., Thomsen, R., Bugai, A., Granneman, S., and Heick Jensen, T. (2023). Temporal-iCLIP captures co-transcriptional RNA-protein interactions. Nat. Commun. *14*, 696. https://doi.org/10.1038/s41467-023-36345-y.

70. Rappsilber, J., Ryder, U., Lamond, A.I., and Mann, M. (2002). Large-scale proteomic analysis of the human spliceosome. Genome Res. *12*, 1231–1245. https://doi.org/10.1101/gr.473902.

71. Makarov, E.M., Owen, N., Bottrill, A., and Makarova, O.V. (2012). Functional mammalian spliceosomal complex E contains SMN complex proteins in addition to U1 and U2 snRNPs. Nucleic Acids Res. *40*, 2639–2652. https://doi.org/10.1093/nar/gkr1056.

72. Sharma, S., Kohlstaedt, L.A., Damianov, A., Rio, D.C., and Black, D.L. (2008). Polypyrimidine tract binding protein controls the transition from exon definition to an intron defined spliceosome. Nat. Struct. Mol. Biol. *15*, 183–191. https://doi.org/10.1038/nsmb.1375.

73. Hsiao, H.-H., Nath, A., Lin, C.-Y., Folta-Stogniew, E.J., Rhoades, E., and Braddock, D.T. (2010). Quantitative characterization of the interactions among c-myc transcriptional regulators FUSE, FBP, and FIR. Biochemistry *49*, 4620–4634. https://doi.org/10.1021/bi9021445.

74. Liu, J., He, L., Collins, I., Ge, H., Libutti, D., Li, J., Egly, J.M., and Levens, D. (2000). The FBP interacting repressor targets TFIIH to inhibit activated transcription. Mol. Cell *5*, 331–341. https://doi.org/10.1016/s1097-2765(00)80428-1.

75. Huang, J.-R., Warner, L.R., Sanchez, C., Gabel, F., Madl, T., Mackereth, C.D., Sattler, M., and Blackledge, M. (2014). Transient electrostatic interactions dominate the conformational equilibrium sampled by multidomain splicing factor U2AF65: a combined NMR and SAXS study. J. Am. Chem. Soc. *136*, 7068–7076. https://doi.org/10.1021/ja502030n.

76. Macias, M.J., Wiesner, S., and Sudol, M. (2002). WW and SH3 domains, two different scaffolds to recognize proline-rich ligands. FEBS Lett. *513*, 30–37. https://doi.org/10.1016/s0014-5793(01)03290-2.

77. Ball, L.J., Kühne, R., Schneider-Mergener, J., and Oschkinat, H. (2005). Recognition of proline-rich motifs by protein-protein-interaction domains. Angew. Chem. Int. Ed. Engl. *44*, 2852–2869. https://doi.org/10.1002/anie.200400618.

78. Zarrinpar, A., Bhattacharyya, R.P., and Lim, W.A. (2003). The structure and function of proline recognition domains. Sci. STKE *2003*, RE8. https://doi.org/10.1126/stke.2003.179.re8.

79. Kofler, M.M., and Freund, C. (2006). The GYF domain. FEBS J. *273*, 245–256. https://doi.org/10.1111/j.1742-4658.2005.05078.x.

80. Sudol, M. (1996). Structure and function of the WW domain. Prog. Biophys. Mol. Biol. *65*, 113–132. https://doi.org/10.1016/s0079-6107(96)00008-9.

81. Mayer, B.J. (2001). SH3 domains: complexity in moderation. J. Cell Sci. *114*, 1253–1263. https://doi.org/10.1242/jcs.114.7.1253.

82. Bell, M.V., Cowper, A.E., Lefranc, M.P., Bell, J.I., and Screaton, G.R. (1998). Influence of intron length on alternative splicing of CD44. Mol. Cell. Biol. *18*, 5930–5941. https://doi.org/10.1128/MCB.18.10.5930.

83. Fox-Walsh, K.L., Dou, Y., Lam, B.J., Hung, S.-P., Baldi, P.F., and Hertel, K.J. (2005). The architecture of pre-mRNAs affects mechanisms of splice-site pairing. Proc. Natl. Acad. Sci. USA *102*, 16176–16181. https://doi.org/10.1073/pnas.0508489102.

84. Dewey, C.N., Rogozin, I.B., and Koonin, E.V. (2006). Compensatory relationship between splice sites and exonic splicing signals depending on the length of vertebrate introns. BMC Genomics *7*, 311. https://doi.org/10.1186/1471-2164-7-311.

85. Gelfman, S., Burstein, D., Penn, O., Savchenko, A., Amit, M., Schwartz, S., Pupko, T., and Ast, G. (2012). Changes in exon-intron structure during vertebrate evolution affect the splicing pattern of exons. Genome Res. *22*, 35–50. https://doi.org/10.1101/gr.119834.110.

86. De Conti, L., Baralle, M., and Buratti, E. (2013). Exon and intron definition in pre-mRNA splicing. Wiley Interdiscip. Rev. RNA *4*, 49–60. https://doi.org/10.1002/wrna.1140.

87. Schneider, M., Will, C.L., Anokhina, M., Tazi, J., Urlaub, H., and Lührmann, R. (2010). Exon definition complexes contain the tri-snRNP and can be directly converted into B-like precatalytic splicing complexes. Mol. Cell *38*, 223–235. https://doi.org/10.1016/j.molcel.2010.02.027.

88. Sharma, S., Wongpalee, S.P., Vashisht, A., Wohlschlegel, J.A., and Black, D.L. (2014). Stem-loop 4 of U1 snRNA is essential for splicing and interacts with the U2 snRNP-specific SF3A1 protein during spliceosome assembly. Genes Dev. *28*, 2518–2531. https://doi.org/10.1101/gad.248625.114.

89. Martelly, W., Fellows, B., Senior, K., Marlowe, T., and Sharma, S. (2019). Identification of a noncanonical RNA binding domain in the U2 snRNP protein SF3A1. RNA *25*, 1509–1521. https://doi.org/10.1261/rna.072256.119.

90. Plaschka, C., Lin, P.-C., Charenton, C., and Nagai, K. (2018). Prespliceosome structure provides insights into spliceosome assembly and regulation. Nature *559*, 419–422. https://doi.org/10.1038/s41586-018-0323-8.

91. Martelly, W., Fellows, B., Kang, P., Vashisht, A., Wohlschlegel, J.A., and Sharma, S. (2021). Synergistic roles for human U1 snRNA stem-loops in pre-mRNA splicing. RNA Biol. *18*, 2576–2593. https://doi.org/10.1080/15476286.2021.1932360.

92. Linares, A.J., Lin, C.-H., Damianov, A., Adams, K.L., Novitch, B.G., and Black, D.L. (2015). The splicing regulator PTBP1 controls the activity of the transcription factor Pbx1 during neuronal differentiation. ELife *4*, e09268. https://doi.org/10.7554/eLife.09268.

93. Delaglio, F., Grzesiek, S., Vuister, G.W., Zhu, G., Pfeifer, J., and Bax, A. (1995). NMRPipe: a multidimensional spectral processing system based on UNIX pipes. J. Biomol. NMR *6*, 277–293. https://doi.org/10.1007/BF00197809.

94. Lee, W., Tonelli, M., and Markley, J.L. (2015). NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy. Bioinformatics *31*, 1325–1327. https://doi.org/10.1093/bioinformatics/btu830.

95. Güntert, P. (2009). Automated structure determination from NMR spectra. Eur. Biophys. J. *38*, 129–143. https://doi.org/10.1007/s00249-008-0367-z.

96. Shen, Y., Delaglio, F., Cornilescu, G., and Bax, A. (2009). TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. J. Biomol. NMR *44*, 213–223. https://doi.org/10.1007/s10858-009-9333-z.

97. Rieping, W., Habeck, M., Bardiaux, B., Bernard, A., Malliavin, T.E., and Nilges, M. (2007). ARIA2: automated NOE assignment and data integration in NMR structure calculation. Bioinformatics *23*, 381–382. https://doi.org/10.1093/bioinformatics/btl589.

98. Laskowski, R.A., Rullmannn, J.A., MacArthur, M.W., Kaptein, R., and Thornton, J.M. (1996). Aqua and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J. Biomol. NMR *8*, 477–486. https://doi.org/10.1007/BF00228148.

99. Bhattacharya, A., Tejero, R., and Montelione, G.T. (2007). Evaluating protein structures determined by structural genomics consortia. Proteins *66*, 778–795. https://doi.org/10.1002/prot.21165.

100. Koradi, R., Billeter, M., and Wüthrich, K. (1996). MOLMOL: A program for display and analysis of macromolecular structures. J. Mol. Graph. *14*, 51–55. https://doi.org/10.1016/0263-7855(96)00009-4.

101. Schrödinger, L., and DeLano, W. (2020). PyMOL. http://www.pymol.org/pymol.

102. Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., et al. (2012). Fiji: an open-source platform for biological-image analysis. Nat. Methods *9*, 676–682. https://doi.org/10.1038/nmeth.2019.

103. Coleman, T., Branch, M.A., and Grace, A. (1999). Optimization Toolbox. For Use with MATLAB. User's guide. The MathWorks Inc, Ver. 2.

104. R Core Team (2016). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. http://www.R-project.org/.

105. Vaquero-Garcia, J., Barrera, A., Gazzara, M.R., González-Vallinas, J., Lahens, N.F., Hogenesch, J.B., Lynch, K.W., and Barash, Y. (2016). A new view of transcriptome complexity and regulation through the lens of local splicing variations. ELife *5*, e11752. https://doi.org/10.7554/eLife.11752.

106. Dosch, J., Bergmann, H., Tran, V., and Ebersberger, I. (2023). FAS: assessing the similarity between proteins using multi-layered feature architectures. Bioinformatics *39*, btad226. https://doi.org/10.1093/bioinformatics/btad226.

107. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21. https://doi.org/10.1093/bioinformatics/bts635.

108. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. *17*, 10–12. https://doi.org/10.14806/ej.17.1.200.

109. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., and Li,

H. (2021). Twelve years of SAMtools and BCFtools. Gigascience *10*, giab008. https://doi.org/10.1093/gigascience/giab008.

110. Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930. https://doi.org/10.1093/bioinformatics/btt656.

111. Roehr, J.T., Dieterich, C., and Reinert, K. (2017). Flexbar 3.0 - SIMD and multicore parallelization. Bioinformatics *33*, 2941–2942. https://doi.org/10.1093/bioinformatics/btx330.

112. Krakau, S., Richard, H., and Marsico, A. (2017). PureCLIP: capturing target-specific protein-RNA interaction footprints from single-nucleotide CLIP-seq data. Genome Biol. *18*, 240. https://doi.org/10.1186/s13059-017-1364-2.

113. Lorenz, R., Bernhart, S.H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., and Hofacker, I.L. (2011). ViennaRNA package 2.0. Algorithms Mol. Biol. *6*, 26. https://doi.org/10.1186/1748-7188-6-26.

114. Huppertz, I., Attig, J., D'Ambrogio, A., Easton, L.E., Sibley, C.R., Sugimoto, Y., Tajnik, M., König, J., and Ule, J. (2014). iCLIP: protein–RNA interactions at nucleotide resolution. Methods *65*, 274–287. https://doi.org/10.1016/j.ymeth.2013.10.011.

115. Spellman, R., Llorian, M., and Smith, C.W.J. (2007). Crossregulation and functional redundancy between the splicing regulator PTB and its paralogs nPTB and ROD1. Mol. Cell *27*, 420–434. https://doi.org/10.1016/j.molcel.2007.06.016.

116. Coelho, M.B., Attig, J., Bellora, N., König, J., Hallegger, M., Kayikci, M., Eyras, E., Ule, J., and Smith, C.W.J. (2015). Nuclear matrix protein Matrin3 regulates alternative splicing and forms overlapping regulatory networks with PTB. EMBO J. *34*, 653–668. https://doi.org/10.15252/embj.201489852.

117. Grzesiek, S., and Bax, A. (1992). Correlating backbone amide and side chain resonances in larger proteins by multiple relayed triple resonance NMR. J. Am. Chem. Soc. *114*, 6291–6293. https://doi.org/10.1021/ja00042a003.

118. Sattler, M., Schleucher, J., and Griesinger, C. (1999). Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. Prog. Nucl. Magn. Reson. Spectrosc. *34*, 93–158. https://doi.org/10.1016/s0079-6565(98)00025-9.

119. Wishart, D.S., and Sykes, B.D. (1994). The 13C chemical-shift index: a simple method for the identification of protein secondary structure using 13C chemical-shift data. J. Biomol. NMR *4*, 171–180. https://doi.org/10.1007/BF00175245.

120. Saitô, H. (1986). Conformation-dependent ^{13}C chemical shifts: a new means of conformational characterization as obtained by high-resolution solid-state ^{13}C NMR. Magn. Reson. Chem. *24*, 835–852. https://doi.org/10.1002/mrc.1260241002.

121. Kjaergaard, M., and Poulsen, F.M. (2011). Sequence correction of random coil chemical shifts: correlation between neighbor correction factors and changes in the Ramachandran distribution. J. Biomol. NMR *50*, 157–165. https://doi.org/10.1007/s10858-011-9508-2.

122. Farrow, N.A., Muhandiram, R., Singer, A.U., Pascal, S.M., Kay, C.M., Gish, G., Shoelson, S.E., Pawson, T., Forman-Kay, J.D., and Kay, L.E. (1994). Backbone dynamics of a free and phosphopeptide-complexed Src homology 2 domain studied by 15N NMR relaxation. Biochemistry *33*, 5984–6003. https://doi.org/10.1021/bi00185a040.

123. Mulder, F.A., Schipper, D., Bott, R., and Boelens, R. (1999). Altered flexibility in the substrate-binding site of related native and engineered high-alkaline *Bacillus subtilisins*. J. Mol. Biol. *292*, 111–123. https://doi.org/10.1006/jmbi.1999.3034.

124. Williamson, M.P. (2013). Using chemical shift perturbation to characterise ligand binding. Prog. Nucl. Magn. Reson. Spectrosc. *73*, 1–16. https://doi.org/10.1016/j.pnmrs.2013.02.001.

125. Zwahlen, C., Gardner, K.H., Sarma, S.P., Horita, D.A., Byrd, R.A., and Kay, L.E. (1998). An NMR experiment for measuring methyl-methyl NOEs in ^{13}C-labeled proteins with high resolution. J. Am. Chem. Soc. *120*, 7617–7625. https://doi.org/10.1021/ja981205z.

126. Marsh, J.A., Singh, V.K., Jia, Z., and Forman-Kay, J.D. (2006). Sensitivity of secondary structure propensities to sequence differences between alpha- and gamma-synuclein: implications for fibrillation. Protein Sci. *15*, 2795–2804. https://doi.org/10.1110/ps.062465306.

127. Linge, J.P., Williams, M.A., Spronk, C.A.E.M., Bonvin, A.M.J.J., and Nilges, M. (2003). Refinement of protein structures in explicit solvent. Proteins *50*, 496–506. https://doi.org/10.1002/prot.10299.

128. Brünger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.S., Kuszewski, J., Nilges, M., Pannu, N.S., et al. (1998). Crystallography & NMR system: a new software suite for macromolecular structure determination. Acta Crystallogr. D Biol. Crystallogr. *54*, 905–921. https://doi.org/10.1107/s0907444998003254.

129. Messias, A.C., and Sattler, M. (2004). Structural basis of single-stranded RNA recognition. Acc. Chem. Res. *37*, 279–287. https://doi.org/10.1021/ar030034m.

130. Wiemann, S., Pennacchio, C., Hu, Y., Hunter, P., Harbers, M., Amiet, A., Bethel, G., Busse, M., Carninci, P., Dunham, I., et al. (2016). The ORFeome Collaboration: a genome-scale human ORF-clone resource. Nature Methods *13*, 191–192.

131. Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M., Sisu, C., Wright, J., Armstrong, J., et al. (2019). GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. *47*. D766–D773. https://doi.org/10.1093/nar/gky955.

132. Busch, A., Brüggemann, M., Ebersberger, S., and Zarnack, K. (2020). iCLIP data analysis: a complete pipeline from sequencing reads to RBP binding sites. Methods *178*, 49–62. https://doi.org/10.1016/j.ymeth.2019.11.008.

133. Paggi, J.M., and Bejerano, G. (2018). A sequence-based, deep learning model accurately predicts RNA splicing branchpoints. RNA *24*, 1647–1658. https://doi.org/10.1261/rna.066290.118.

134. Hinrichs, A.S., Karolchik, D., Baertsch, R., Barber, G.P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T.S., Harte, R.A., Hsu, F., et al. (2006). The UCSC genome browser database: update 2006. Nucleic Acids Res. *34*. D590–D598. https://doi.org/10.1093/nar/gkj144.

135. Green, C.J., Gazzara, M.R., and Barash, Y. (2018). MAJIQ-SPEL: web-tool to interrogate classical and complex splicing variations from RNA-Seq data. Bioinformatics *34*, 300–302. https://doi.org/10.1093/bioinformatics/btx565.

136. Norton, S.S., Vaquero-Garcia, J., Lahens, N.F., Grant, G.R., and Barash, Y. (2018). Outlier detection for improved differential splicing quantification from RNA-Seq experiments with replicates. Bioinformatics *34*, 1488–1497. https://doi.org/10.1093/bioinformatics/btx790.

137. Zhang, J., Bajari, R., Andric, D., Gerthoffert, F., Lepsa, A., Nahal-Bose, H., Stein, L.D., and Ferretti, V. (2019). The International Cancer Genome Consortium data portal. Nat. Biotechnol. *37*, 367–369. https://doi.org/10.1038/s41587-019-0055-9.

138. Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E., et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. *2*, 401–404. https://doi.org/10.1158/2159-8290.CD-12-0095.

139. Gao, J., Aksoy, B.A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S.O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci. Signal. *6*, pl1. https://doi.org/10.1126/scisignal.2004088.

140. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P.,

et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443. https://doi.org/10.1038/s41586-020-2308-7.

141. Tate, J.G., Bamford, S., Jubb, H.C., Sondka, Z., Beare, D.M., Bindal, N., Boutselakis, H., Cole, C.G., Creatore, C., Dawson, E., et al. (2019). COSMIC: the catalogue of somatic mutations in cancer. Nucleic Acids Res. *47*. D941–D947. https://doi.org/10.1093/nar/gky1015.

142. Grossman, R.L., Heath, A.P., Ferretti, V., Varmus, H.E., Lowy, D.R., Kibbe, W.A., and Staudt, L.M. (2016). Toward a shared vision for cancer genomic data. N. Engl. J. Med. *375*, 1109–1112. https://doi.org/10.1056/NEJMp1607591.

143. Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., et al. (2018). ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. *46*. D1062–D1067. https://doi.org/10.1093/nar/gkx1153.

144. Yeo, G., and Burge, C.B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J. Comput. Biol. *11*, 377–394. https://doi.org/10.1089/1066527041410418.

145. Birikmen, M., Bohnsack, K.E., Tran, V., Somayaji, S., Bohnsack, M.T., and Ebersberger, I. (2021). Tracing eukaryotic ribosome biogenesis factors into the archaeal domain sheds light on the evolution of functional complexity. Front. Microbiol. *12*, 739000. https://doi.org/10.3389/fmicb.2021.739000.

146. Smith, T., Heger, A., and Sudbery, I. (2017). UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Res. *27*, 491–499. https://doi.org/10.1101/gr.209601.116.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Rabbit anti-FUBP1	GeneTex	Cat# GTX104579; RRID: AB_11165485
Mouse anti-U2AF2	Sigma-Aldrich	Cat# U4758; RRID: AB_262122
Mouse anti-SF3B1	MBL	Cat# D221-3; RRID: AB_592712
Mouse anti-SF1	Abnova	Cat# H00007536-M01A; RRID: AB_10774630
rabbit anti-PTBP1	Christopher Smith	Linares et al.[92]
Mouse anti-vinculin	Sigma-Aldrich	Cat# V9264; RRID: AB_10603627
Goat anti-rabbit IgG, HRP-linked	Cell Signaling	Cat# 7074S; RRID: AB_2099233
Horse anti-mouse IgG, HRP-linked	Cell Signaling	Cat# 7076S; RRID: AB_330924
Bacterial and virus strains		
DH5alpha	Invitrogen	Cat# 18265017
MACH1	Invitrogen	Cat# C862003
E. coli BL21-CodonPlus (DE3)-RIL	Agilent	Cat# 230245
E. coli BL21 (DE3)	Sigma-Aldrich	Cat# CMC0014
Chemicals, peptides, and recombinant proteins		
FUGENE HD reagent	Promega	Cat# E2311
Lipofectamine CRISPRMAX reagent	Thermo Fisher	Cat# CMAX00001
Lipofectamine RNAimax	Thermo Fisher	Cat# 13778150
Lipofectamine 2000	Invitrogen	Cat# 11668019
cOmplete Protease-Inhibitor Mix	Sigma-Aldrich	Cat# 4693159001
TURBO DNase	Thermo Fisher	Cat# AM2238
SuperSignal West PICO Chemiluminescent Substrate	Thermo Fisher	Cat# 15626144
4-thiouridine	Sigma-Aldrich	Cat# T4509-25MG
T4 RNA ligase	New England Biolabs	Cat# M0202S
T4 RNA ligase 1	New England Biolabs	Cat# M0437M
pCp-Cy5	Jena Bioscience	Cat# NU-1706-CY5
T7 RNA polymerase	Geerlof A., Protein Expression and Purification Facility, HMGU Munich	N/A
Pfu DNA Polymerase	Promega	Cat# M7741
OneTaq DNA Polymerase	New England Biolabs	Cat# M0480S
Phusion High-Fidelity DNA Polymerase	New England Biolabs	Cat# M0530S
Critical commercial assays		
TranscriptAid Enzyme Mix	Thermo Fisher	Cat# K0441
GeneArt Genomic Cleavage Detection Assay	Thermo Fisher	Cat# A24372
Zero Blunt TOPO PCR Cloning Kit	Thermo Fisher	Cat# 451245
RNeasy PLUS Mini Kit	Qiagen	Cat# 74034
TruSeq library preparation Kit “Ribo-Zero Gold”	Illumina	Cat# 20040526
RevertAid First Strand cDNA Synthesis Kit	Thermo Fisher	Cat# 10161310
Q5 Site-Directed Mutagenesis Kit	New England Biolabs	Cat# E0552S
High Sensitivity D1000 ScreenTape	Agilent	Cat# 5067-5584
High Sensitivity RNA ScreenTape	Agilent	Cat# 5067-5579
NuPAGE 1 mm, 4-12% Bis-Tris Mini Protein Gel	Thermo FIsher	Cat# 12090156
HiScribe T7 High Yield RNA Synthesis Kit	New England Biolabs	Cat# E2040S

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
ProNex Dual Size-Selective Purification System	Promega	Cat# NG2002
BP clonase II mix kit	Invitrogen	Cat# 10348582
LR clonase technology	Invitrogen	Cat# 11791020
Deposited data		
in vitro and *in vivo* iCLIP and RNA-Seq data	This study	GEO: GSE220186
Kinetic modeling of cassette exon splicing	This study	https://doi.org/10.5281/zenodo.8076768
Protein structure data	This study	PDB: 8P25
NMR data	This study	BMRB: 34816
Original Western blot, gel images and capillary electrophoresis images	This study, Mendeley Data	https://doi.org/10.17632/nj8ybm8vb2.1
RNA-Seq data: control and shRNA knockdown for FUBP1 in K562 cells	Luo et al.[63], ENCODE Project Consortium[62]	ENCODE: ENCSR260BQC (control) and ENCSR608IXR (FUBP1 KD)
Differentially spliced junctions in splicing factor mutations	Seiler et al.[1]	Table S3 in Seiler et al.
Experimental models: Cell lines		
human: HeLa	ATCC	Cat# CCL-2, RRID:CVCL_0030
human: RPE1 FUBP1 WT: hTERT-RPE1 NatNeo Cas9 Mono Puro sens	Manuel Kaulich	N/A
human: RPE1 FUBP1 KO: hTERT-RPE1 NatNeo Cas9 Mono Puro sens FUBP1 -/-	This study	N/A
human: RPE1 FUBP1 Nbox-mut: hTERT-RPE1 NatNeo Cas9 Mono Puro sens FUBP1 indel 31-40	This study	N/A
human: HEK293	DSMZ	ACC305
Oligonucleotides		
See Table S5	(too many oligos to list here)	N/A
Recombinant DNA		
See Table S6	(too many plasmids to list here)	N/A
Software and algorithms		
Topspin 3.5	Bruker	https://www.bruker.com/en/products-and-solutions/mr/nmr-software/topspin.html
NMRpipe	Delaglio et al.[93]	https://www.ibbr.umd.edu/nmrpipe/index.html
NMRFAM-Sparky	Lee et al.[94]	https://nmrfam.wisc.edu/nmrfam-sparky-distribution/
CYANA 3.98.13	Güntert[95]	https://cyana.org/wiki/Main_Page
TALOS+	Shen et al.[96]	https://spin.niddk.nih.gov/bax/software/TALOS/
ARIA2.3	Rieping et al.[97]	http://aria.pasteur.fr/
ProcheckNMR	Laskowski et al.[98]	https://www.ebi.ac.uk/thornton-srv/software/PROCHECK/
PSVS	Bhattacharya et al.[99]	https://montelionelab.chem.rpi.edu/PSVS/PSVS/
MolMol	Koradi et al.[100]	https://sourceforge.net/p/molmol/wiki/Home/
PYMOL	Schrödinger and DeLano[101]	https://pymol.org/2/
ImageJ 2.1.0	Schindelin et al.[102]	https://imagej.net/
MicroCalPEAQ ITC Analysis software	Malvern Panalytical	https://www.malvernpanalytical.com/
Agilent TapeStation Software 5.1	Agilent	https://www.agilent.com
Image Lab 6.0.1 build 34	bio-rad	https://www.bio-rad.com/
MATLAB	Coleman et al.[103]	https://www.mathworks.com/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
R 4.1.1.	Core Team[104]	https://www.r-project.org/
MAJIQ v2.3	Vaquero-Garcia et al.[105]	https://majiq.biociphers.org/
FAS	Dosch et al.[106]	https://github.com/BIONF/FAS
fDOG	N/A	https://github.com/BIONF/fDOG
STAR	Dobin et al.[107]	https://github.com/alexdobin/STAR
Cutadapt 2.4	Martin[108]	https://cutadapt.readthedocs.io/en/stable/
Samtools v1.9	Danecek et al.[109]	http://www.htslib.org/
Subread tool suite v1.6.2	Liao et al.[110]	https://subread.sourceforge.net/
FastQC v0.11.8	N/A	https://www.bioinformatics.babraham.ac.uk/projects/fastqc
FASTX-Toolkit v0.0.14	N/A	http://hannonlab.cshl.edu/fastx_toolkit/
seqtk v1.3	N/A	https://github.com/lh3/seqtk/
Flexbar v3.4.0	Roehr et al.[111]	https://github.com/seqan/flexbar
PureCLIP v1.3.1	Krakau et al.[112]	https://github.com/skrakau/PureCLIP
ViennaRNA Package 2.4.17	Lorenz et al.[113]	https://www.tbi.univie.ac.at/RNA/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Julian König (j.koenig@imb-mainz.de).

Materials availability

All unique/stable reagents generated in this study are available from the lead contact.

Data and code availability

- RNA-seq, *in vivo* and *in vitro* iCLIP data have been deposited at GEO and are publicly available as of the date of publication. Accession numbers are listed in the key resources table. Protein structures have been deposited to the Protein Data Bank and are available under the accession number 8P25. NMR data used for structure calculation are deposited in the BMRB under the accession code 34816. Original Western blot, gel images and capillary electrophoresis images have been deposited at Mendeley Data and are publicly available as of the date of publication. The DOI is listed in the key resources table.
- This paper analyses existing, publicly available data. These accession numbers for the datasets are listed in the key resources table.
- All original code has been deposited at GitHub and is publicly available as of the date of publication at https://doi.org/10.5281/zenodo.8076768.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

RPE1 cell lines and culture conditions

The hTERT-RPE1 NatNeo Cas9 Mono Puro sens cell line was a generous gift of the Kaulich lab at the Frankfurt CRISPR/Cas Screening Center (FCSC) and are modified from original hTERT RPE1 cells (ATCC, CRL-4000). Cells were grown and maintained in Dulbecco's modified Eagle's medium (DMEM): Nutrient Mixture F-12 (DMEM/F-12; Thermo Fisher 11530566), supplemented with 10% fetal bovine serum (PAN-Biotech), 2 mM glutamine (Thermo Fisher), 1% penicillin–streptomycin (Thermo Fisher), and 20 μg/ml hygromycin B (Thermo Fisher). Cells were incubated at 37°C with 5% CO_2. Subcultivation was performed with 3 ml of 0.1% trypsin every 2–3 days for 20 passages. After that, new cells were thawed from stocks containing 1×10^6 cells in 1 ml of growth medium, supplemented with 10% DMSO and 50% fetal bovine serum (FBS). For semi-quantitative RT-PCR, 1×10^5 RPE1 cells were seeded into one well of a six-well plate (Falcon), one day prior to transfection. DNA (2 μg) was diluted in 100 μl of OptiMEM and transfected with 6.4 μl of Fugene HD reagent (Promega). Cells were incubated at 37°C with 5% CO_2 for 24 h before harvesting. For RNA-seq, 1.5×10^6 cells were seeded in a 10-cm cell culture dish (Corning) 48 h prior to isolation.

HeLa cell line and culture conditions

HeLa cells (ATCC CCL-2) were grown and maintained in DMEM (Thermo Fisher), supplemented with 10% FBS, 2 mM glutamine (Thermo Fisher) and 1% penicillin–streptomycin (Thermo Fisher). Cells were incubated at 37°C with 5% CO_2. Subcultivation was performed with 3 ml of 0.1% trypsin every 2–3 days for 20 passages. After that, new cells were thawed from stocks containing 1×10^6 cells in 1 ml of growth medium, supplemented with 10% DMSO (Sigma) and 50% FBS.

HEK cell line and culture conditions

HEK293 cells (DSMZ) were grown and maintained in DMEM (Thermo Fisher), supplemented with 10% fetal bovine serum (PAN-Biotech), 2 mM glutamine (Thermo Fisher) and 1% penicillin–streptomycin (Thermo Fisher). Cells were incubated at 37°C with 5% CO_2. Subcultivation was performed with 1 ml of 0.05% trypsin every 2–3 days for up to 15 passages. Then, new cells were thawed from stocks containing 2×10^6 cells in 1 ml of growth medium, supplemented with 10% DMSO (Sigma) and 90% FBS.

Recombinant protein expression

Proteins were expressed in *E. coli* BL21 (DE3) cells grown in LB medium or M9 minimal medium supplemented with 1 g/l $^{15}NH_4Cl$ and 2 g/l ^{13}C-glucose (uniformly labeled) at 37°C. Protein expression was induced with 1.0 mM isopropyl β-D-1-thiogalactopyranoside (IPTG).

METHOD DETAILS

Establishing *FUBP1* KO/*Nbox*^{mut} cell lines

FUBP1 was mutated and knocked out using the CRISPR/Cas9 system in hTERT-RPE1 NatNeo mono puro sens cells. This cell line is puromycin sensitive and expresses *Streptococcus pyogenes* Cas9 under neomycin resistance. For the creation of the *FUBP1* KO and *FUBP1-Nbox*^{mut} RPE1 cell lines, cells were cultured as described above with the addition of neomycin (G418, InvivoGen) to preserve Cas9 expression. Guide RNA (gRNA) was amplified from oligos #54 and #55 (Table S5) with Phusion Polymerase (New England Biolabs) and *in vitro* transcribed with TranscriptAid Enzyme Mix (Thermo Fisher) according to the manufacturer's protocol. Cells were then transfected with the resulting gRNA using Lipofectamine CRISPRMAX (Thermo Fisher) according to the manufacturer's protocol and incubated for 48 h. To assess the general editing efficiency, a GeneArt Genomic Cleavage Detection Assay (Thermo Fisher) was performed. Edited cells were then sorted by fluorescence-activated cell sorting (FACS), and each cell was cultured in a separate well of a 96-well plate (Corning). From each clonal cell line, genomic DNA (gDNA) was isolated and amplified by PCR. The successful disruption of the targeted site was validated by enzyme restriction and Sanger sequencing (StarSEQ GmbH, Mainz, Germany) of the colonies. To obtain the novel sequence of the targeted site on both alleles, gDNA was also cloned into TOPO vectors using the Zero Blunt TOPO PCR Cloning Kit (Thermo Fisher), and the obtained plasmids were Sanger-sequenced. All Sanger sequencings were performed with oligo #56 (Table S5). The edited sequences led to mutated protein products, as shown in Figure S5G.

Immunoblotting

For each hTERT RPE1-derived cell line, 1×10^6 cells were seeded on a 10-cm cell culture dish (Corning) and harvested after incubation for 48 h at 37°C, 5% CO_2. Cells were lysed in modified RIPA buffer containing 50 mM Tris-HCl, 150 mM NaCl, 1 mM EDTA, 1% NP-40 (Sigma), 0.1% sodium deoxycholate (Sigma) and supplemented with cOmplete Protease Inhibitor Mix (Sigma), and TURBO DNase (Thermo Fisher) for 15 min on ice. Cell debris was precipitated by centrifugation at 16,000 ×*g* for 15 min at 4°C. The cleared protein lysate was transferred into a new reaction tube (Eppendorf) and the concentration was measured with a BCA Protein Assay Kit (Thermo Fisher). 20 μg of protein lysate was mixed with 4× NuPAGE LDS Sample Buffer and heated to 70°C for 10 min. Samples were loaded onto a NuPAGE 1 mm, 4–12% Bis-Tris Mini Protein Gel (Thermo Fisher) and electrophoresis was performed at 180 V, 400 mA for 50 min on a NuPAGE Novex Gel System (Invitrogen). Protein transfer to a nitrocellulose membrane (VWR International) was performed at 30 V, 400 mA over 60 min using the same gel system. The membrane was blocked in 5% milk diluted in PBS-T. The primary antibody (key resources table) was incubated overnight at 4°C, and the secondary antibody was incubated for 60 min at room temperature. All antibodies were diluted in 5% milk–PBS-T. Between blocking and primary and secondary antibody steps, the membrane was washed three times with PBS-T. Detection was performed with SuperSignal West PICO Chemiluminescent Substrate (Thermo Fisher) and BioRad GelDoc (BioRad).

RPE1 RNA-seq

For RPE1 RNA sequencing (ID: imb_koenig_2020_12) and semi-quantitative RT-PCR analysis, RPE1 cells were grown as described above. Cells were washed once with DPBS and harvested with a cell scraper in l ml of DPBS. Suspensions were centrifuged at 1,000 ×*g* for 1 min at 4°C. RNA was isolated from cell pellets using an RNeasy PLUS Mini Kit (Qiagen) according to the manufacturer's protocol. For sequencing, RNA concentration was measured by Qubit RNA BR Assay and integrity of the RNA was confirmed by Bioanalyzer RNA Nano Assay (Agilent). Ribosomal RNA was removed and the remaining RNA was reverse transcribed into cDNA using the TruSeq library preparation kit with Ribo-Zero Gold (Illumina). The libraries were sequenced on an Illumina NextSeq 500 sequencer as 159-nt single-end reads.

HeLa RNA-seq

200,000 cells were seeded per well in a six-well dish 24 h prior to siRNA treatment. RNA-seq to assess intron splicing in HeLa cells (ID: imb_koenig_2018_18) was performed in four replicates. HeLa cells underwent a control knockdown (KD) with no-target siRNA. Oligos #40–#43 (Table S5) were delivered into cells using 3 μl of Lipofectamine RNAimax (Thermo Fisher) in 100 μl of OptiMEM to achieve a final siRNA concentration of 20 nM. Cells were harvested after incubation for 48 h. RNA was isolated from cell pellets using RNeasy PLUS Mini Kit (Qiagen) according to the manufacturer's protocol. RNA concentration was measured by Qubit RNA BR Assay, and integrity of the RNA was confirmed by Bioanalyzer RNA Nano Assay (Agilent). Ribosomal RNA was removed and the remaining RNA was reverse transcribed into cDNA using the TruSeq library preparation kit with Ribo-Zero Gold (Illumina). RNA-seq samples were sequenced on an Illumina NextSeq 500 sequencer as 84-nt single-end reads.

Semi-quantitative RT-PCR

The *MPDZ* minigene was created from HeLa gDNA extracts by amplification of chr9:13,183,353-13,189,041 with Phusion HighFidelity Polymerase (New England Biolabs). The PCR fragment was cloned into a pCR2.1 vector by Gibson assembly (IMB Protein Production Core Facility). *MPDZ* introns were shortened using a Q5 Site-Directed Mutagenesis Kit (New England Biolabs), resulting in $MPDZ^{\Delta intron}$, which lacks chr9:13,186,637-13,188,633 and chr9:13,183,736-13,186,120, $MPDZ^{\Delta BS}$, lacking chr9:13,186,494-13,186,618 and chr9:13,183,632-13,186,718, and $MPDZ^{\Delta intron+\Delta BS}$, lacking chr9:13,186,494-13,188,633 and chr9:13,183,632-13,186,120 (Figure S6B). The open reading frames for GFP and the FUBP1 variants ($FUBP1^{FL}$, $FUBP1^{\Delta N}$, $FUBP1^{A38D}$, $FUBP1^{\Delta C}$, $FUBP1^{W586,615R}$) used in the complementation assay were integrated in a pcDNA5 vector containing a CMV promoter and an N-terminal GFP tag, which was then used to transform DH5alpha cells (Invitrogen). All expression vectors and minigenes are described in Table S6. Plasmid purification was performed with the Qiaprep Spin Miniprep Kit (Qiagen) or the Qiaprep Plasmid Plus Midi Kit (Qiagen). Sequences were verified by Sanger sequencing. All hTERT RPE1 cell lines were seeded, transfected, and harvested as described in the section "RPE1 cell culture". For complementation, an equimolar amount of expression vector and minigene was used. RNA was isolated with the RNeasy Plus Mini Kit (Qiagen) and reverse transcribed using the RevertAid First Strand cDNA Synthesis Kit (Thermo Fisher). The minigene cDNA was then amplified using OneTaq DNA Polymerase according to the manufacturer's protocol and oligos #57 and #58 as primers (Table S5). Splicing products were assessed on a High Sensitivity D1000 ScreenTape (Agilent) (Figure S6D). The percent spliced-in (PSI) value for the alternative exon was determined using the following formula: *Inclusion / (Inclusion + Skipping)*. PSI values in the complementation experiment were normalized to the mean of the wild-type (WT) within each condition. Statistical significance was assessed by Student's t-test and multiple testing correction was performed using the false discovery rate (FDR).

In vivo iCLIP

In vivo iCLIP was used to study protein–RNA interactions with individual nucleotide resolution.[43] For the U2AF2 *in vivo* iCLIP study, data from two iCLIP experiments were combined. The first U2AF2 and PTBP1 *in vivo* iCLIP experiments were performed as previously described.[114] The second U2AF2 *in vivo* iCLIP experiment as well as *in vivo* iCLIP experiments on FUBP1, SF1, and SF3B1 were performed using the iCLIP2 protocol as previously described.[44] In brief, HeLa cells were irradiated (150 mJ/cm^2) in a CL1000 UV crosslinker (UPV) to covalently bond the RNA-binding proteins to the bound nucleic acids. For *in vivo* iCLIP of FUBP1, crosslinking was achieved by 4-thiouridine (4sU)-mediated crosslinking (see section below). During subsequent cell lysis, the lysate was DNase-treated with TURBO DNase (Thermo Fisher) and RNA was partially digested to create 50–200-nt fragments. Immunoprecipitation of the investigated proteins was performed with antibodies listed in the key resources table. The anti-PTBP1 antibody was a kind gift from Christopher Smith.[115] Radioactive labeling at the 3′ end of the precipitated RNA enables visualization of the RNP complex by SDS-PAGE and transfer to a nitrocellulose membrane. After recovery of protein–RNA complexes from the membrane, proteinase K digestion resulted in protein-free RNA. cDNA was synthesized by reverse transcription, which stops at the crosslinked site, leading to truncated reads in the sequencing. The cDNA was cleaned twice using MyONE Silane beads (Thermo Fisher). PCR amplification and ProNex size selection were performed to amplify and purify the library, respectively. *In vivo* iCLIP libraries (except PTBP1 libraries) were sequenced on an Illumina NextSeq 500 sequencer as 92-nt single-end reads including a 6-nt (or 4-nt in the case of the first U2AF2 iCLIP) sample barcode as well as 5+4-nt (or 3+2-nt) unique molecular identifiers (UMIs). PTBP1 iCLIP libraries were sequenced on an Illumina GA-II machine[116] and then re-sequenced on an Illumina HiSeq 2000 machine as 50-nt single-end reads including a 4-nt sample barcode and 3+2-nt UMIs.

4-thiouridine crosslinking of FUBP1 in vivo iCLIP

For the FUBP1 *in vivo* iCLIP, HeLa cells were 4sU-labeled by adding 0.1 M 4sU in DMSO to a final concentration of 100 μM in a 10-cm cell culture dish. Cells were incubated for 16 h at 37°C, 5% CO_2, with the exclusion of light. After incubation, the cells were moved onto ice, shielded from light and irradiated at 365 nm, 800 mJ. Then, iCLIP was performed as described above.

In vitro iCLIP

In vitro iCLIP measures the intrinsic RNA-binding affinity of an RNA-binding protein (RBP).[28] To that end, recombinant proteins and *in vitro* transcripts resembling long natural transcripts[28] or a large-scale RNA pool transcribed from an oligonucleotide library were mixed and subjected to UV crosslinking and immunoprecipitation of the RBP of interest.

Production of recombinant proteins

N-terminally 6xHis-tagged U2AF2^{RRM12} was purified as previously described.[28] In brief, a recombinant construct (Table S6) was expressed in *E. coli* BL21-CodonPlus (DE3)-RIL cells (Agilent) for 3–4 h at 37°C using LB-Media and 1 mM IPTG. U2AF2^{RRM12} was purified using Ni Sepharose 6 Fast Flow beads (GE Healthcare) according to the manufacturer's protocol, and concentrated with Spin-X UF 500 5K MWCO columns (Corning) to a concentration of 1.156 mg/ml before being flash-frozen in liquid nitrogen and stored at −80°C. All three N-terminally 6xHis-tagged FUBP1 protein variants (FUBP1^{FL}, FUBP1$^{\Delta N}$, FUBP1^{N74}; Table S6) were expressed overnight at 16°C using LB media and 1 mM IPTG. Cells were lysed in lysis buffer (50 mM Tris-Cl, pH 8.0, 500 mM NaCl, 1 mM DTT, 5% glycerol, EDTA-free cOmplete protease inhibitor cocktail), using a CF1 Cell Disrupter (Constant Systems). Lysates were cleared by centrifugation (40,000 ×*g*, 30 min, 4°C). Recombinant proteins were affinity-purified from cleared lysates using an NGC Quest Plus FPLC system (Biorad) and a HisTrap FF 5 ml column (Cytiva) according to the manufacturers' protocols. Full-length FUBP1^{FL} and FUBP1$^{\Delta N}$ proteins were diluted 1:10 in heparin binding buffer (30 mM Na-HEPES, 20 mM NaCl, 5% glycerol, 1 mM DTT, pH 7.4), loaded onto a Heparin HP 5 ml column (Cytiva) and eluted over 15 column volumes using a linear gradient of 20–1000 mM NaCl in the heparin binding buffer. All FUBP1 variants were concentrated using Amicon 15 ml spin concentrators (Merck Millipore) and subjected to gel filtration (Superdex 200 16/60 pg in 30 mM Na-HEPES, 100 mM NaCl, 1 mM DTT, 5% glycerol, pH 7.4). Peak fractions containing the recombinant proteins after gel filtration were pooled, and protein concentration was determined by using absorbance spectroscopy and the respective extinction coefficient at 280 nm, before aliquots were flash-frozen in liquid nitrogen and stored at −80°C. For the detailed workflow, log files can be requested from Dr. Julian König.

Preparation of *long* in vitro transcripts

Long *in vitro* transcripts were prepared as described in Sutandy et al.[28] Minigene and spike-in RNAs were created by PCR amplification of DNA templates using Phusion High-Fidelity DNA Polymerase (New England Biolabs) according to the manufacturer's protocol. *In vitro* transcription of gel-purified PCR products was performed using HiScribe T7 High Yield RNA Synthesis Kit (New England Biolabs) according to the manufacturer's instructions. RNA was isolated with the RNeasy Plus Mini Kit (Qiagen), followed by DNA digestion with TURBO DNase and another RNA extraction. RNA quality was verified by capillary electrophoresis using High Sensitivity RNA ScreenTape (Agilent). RNA concentration was measured with a Qubit RNA HS Assay Kit (Thermo Fisher). Aliquots of equimolar mixes of all minigenes as well as spike-in aliquots were stored at −80°C.

In vitro *iCLIP with long* in vitro *transcripts*

In vitro iCLIP with long *in vitro* transcripts (ID: imb_koenig_2018_01_sub16) was performed for U2AF2^{RRM12} alone or supplemented with different FUBP1 variants. The experiment was performed with a pool of eight *in vitro* transcripts (*C4BPB*, *MPDZ*, *MYC*, *MYL6*, *NF1*, *TENT2*, *PCBP2*, and *PTBP2*, see GEO record GSE220183) as previously described.[28] The *in vitro* transcripts were preheated for 5 min at 70°C to minimize RNA secondary structure. Then, *in vitro* transcripts at a final concentration of 2 nM were added to 50 nM U2AF2^{RRM12} either alone (three replicates) or supplemented with either 50 nM FUBP1^{FL} (two replicates), 50 nM FUBP1$^{\Delta N}$ (two replicates), or 50 nM FUBP1^{N74} (two replicates). The mixtures were incubated at 37°C for 5 min before UV irradiation at 50 mJ/cm^2. The *in vitro* iCLIP reaction was spiked with 10 μl of crosslinked mixture containing 250 nM U2AF2^{RRM12} and 6 nM *NUP133 in vitro* transcript for normalization.[28] Partial RNase digestion and DNase treatment, followed by the standard iCLIP protocol, were performed as described in the section "*In vivo iCLIP*". After reverse transcription, the cDNA was purified and libraries were generated according to the iCLIP2 protocol.[44]

Preparation of oligo-derived transcripts

A total of 1,998 DNA oligonucleotides were chosen to represent 182-nt regions around 3′ splice sites, including the last 132 nt upstream of a 3′ splice site and the first 50 nt of the downstream exon, preceded by 18 nt of T7 promoter sequence for the reverse transcription. The genomic coordinates of all regions represented in the oligonucleotide library are listed in GEO record GSE220183. The DNA oligonucleotides were purchased from TWIST Bioscience (South San Francisco, CA). Before *in vitro* transcription, L3 adapter ligation was performed. This was achieved by resuspending the DNA pellet in T4 RNA ligase (New England Biolabs) mix containing a 1:10 oligo/adapter ratio for high ligation efficiency. This mixture was reacted overnight at 16°C at 1300 rpm and then inactivated at 98°C for 5 min. L3-ligated DNA oligonucleotide (2.6 ng) was amplified using the Phusion High-Fidelity DNA Polymerase (New England Biolabs) according to the manufacturer's protocol. Amplicons were purified twice using the ProNex Dual Size-Selective Purification System (Promega) with an optimized bead/library ratio of first 1.13 and then 0.5. Capillary electrophoresis with a High Sensitivity D1000 ScreenTape (Agilent) was used for quality control. Then, *in vitro* transcription was performed for 4 h at 37°C by following the HiScribe T7 (New England Biolabs) protocol for short transcripts. Subsequently, RNA was treated with TURBO DNase I and isolated using Qiagen's protocol for "Total RNA containing small RNA from cells" (RNeasy Plus Mini Handbook, Appendix E) with the reagents mentioned above.

in vitro *iCLIP on oligo-derived transcripts*

For *in vitro* iCLIP with an oligonucleotide-derived RNA pool (ID: imb_koenig_2018_01_sub12), the oligonucleotide-derived transcript pool at a concentration of 50 nM was preheated for 5 min at 70°C and incubated with 50 nM U2AF2^{RRM12} alone or with either 50 or 300 nM FUBP1^{FL} (three replicates each) for 10 min before UV irradiation at 50 mJ/cm^2. iCLIP was performed as described in the section "*In vivo iCLIP*", omitting the partial RNase digestion and L3 linker ligation steps as they do not apply here. The reaction was spiked with a mix of 10 150-nt long spike-in oligonucleotides for normalization (oligos #44–#53; Table S5).

Sequencing and data preprocessing

In vitro iCLIP libraries were sequenced on an Illumina NextSeq 500 sequencer as 150-nt single-end reads including a 6-nt sample barcode as well as 5+4-nt UMIs. The reads were bioinformatically preprocessed as described for *in vivo* iCLIP samples. The number of uniquely mapped reads for all *in vitro* iCLIP samples are given in Table S1.

Protein expression and purification

All plasmids encoding sequences of FUBP1, U2AF2, chimeric U2AF2$^{linker\text{-}RRM2}$/FUBP1$^{N\text{-}box}$ (linked by a 14 GS linker), SF1, SNRPA, SNRPB, and PRPF40B were cloned into the pETM11 vector or pET24 vector with a His tag, His-GB1 tag, or His-protein A tag, followed by a TEV cleavage site. The point mutants of FUBP1 were generated by site-directed mutagenesis. All constructs are listed in Table S6.

Recombinant proteins were expressed in *E. coli* BL21 (DE3) cells in LB medium or M9 minimal medium supplemented with 1 g/l $^{15}NH_4Cl$ and 2 g/l ^{13}C-glucose (uniformly labeled). After growth of the bacterial cells to an OD_{600} value of 0.8, protein expression was induced with 1.0 mM IPTG followed by overnight expression at 18°C. After resuspension in 50 mM Tris, pH 8.0, 500 mM NaCl, 10 mM imidazole (supplemented with lysozyme, 1 mg/ml DNase, 2 mM $MgSO_4$, and protease inhibitor), the cells were lysed using a French press. Cleared lysates were added to Ni–NTA resin, washed with 2 M NaCl and eluted with 500 mM imidazole. The His tag was cleaved with His-tagged TEV protease at 4°C overnight. The protein was further purified by removing the cleaved His tag, uncleaved protein and TEV protease from the desired protein on a second Ni–NTA column. All proteins were further purified by ion-exchange chromatography on RESOURCE S or RESOURCE Q columns (Cytiva) (20 mM Tris, pH 8.0 or 20 mM sodium phosphate, pH 6.5, gradient from 0 to 1 M NaCl in 10 column volumes) followed by size-exclusion chromatography on a HiLoad 16/600 Superdex 75 column (GE Healthcare) (20 mM sodium phosphate, pH 6.5, 150 mM NaCl).

NMR spectroscopy

All NMR samples ($^{13}C^{15}N$- or ^{15}N-labeled, as appropriate) were measured at concentrations of 0.1–1 mM in NMR buffer (20 mM sodium phosphate, pH 6.5, 50 mM NaCl, 2 mM DTT) containing 10% (v/v) D_2O at 25°C on 900-, 800-, 600-, or 500-MHz Bruker Avance NMR spectrometers (cryogenic triple-resonance gradient probes). The NMR spectra were processed with TOPSPIN3.5 (Bruker) or NMRPipe[93] and analyzed using NMRFAM-Sparky.[94]

Chemical shift assignment

Protein backbone assignments were obtained from standard HNCA, HNCACB, CBCA(CO)NH, HNHA backbone experiments. Specifically, for KH domains, the ^1H–^{15}N HSQC spectrum of KH1–4 was first assigned, then corresponding assignments were transferred to the spectra of the individual and tandem KH domains. Further side-chain resonances were assigned using CC(CO)NH, HCC(CO)NH, hCCH-TOCSY and HcCH-TOCSY experiments. The distance restraints for structure calculations were obtained from 3D ^{15}N- and ^{13}C-edited NOESY–HSQC experiments.[117,118] Secondary structure propensities were derived from the difference of C_a and C_b chemical shifts to the random coil shifts.[119–121]

Relaxation experiments

^{15}N-relaxation experiments were recorded on an 800 MHz Bruker Avance NMR spectrometer at 25°C and ^{15}N T_1 and T_2 relaxation times were acquired from pseudo-3D HSQC experiments in an interleaved manner with eight relaxation delays for T_1 (20, 60, 100, 200, 400, 600, 800, 1200 ms) and nine relaxation delays for T_2 (16.96, 33.92, 67.84, 101.76, 135.68, 169.6, 254.4, 305.28, 339.2 ms).[122] Residual relaxation rates were obtained by fitting the data to an exponential function using NMRFAM-Sparky.[94]

Titrations

For NMR titrations, ^1H–^{15}N HSQC spectra were measured after each addition of titrant and the changes were visualized by calculating the CSP.[123] The K_D values were calculated from NMR titrations by plotting the CSP of selected peaks (8) against the ligand concentration and fitting the data as previously described. Standard deviations of the mean were calculated from K_D values of the 8 selected peaks.[124]

Structure calculation

To stabilize the U2AF2 and FUBP1 interaction, a chimeric construct of U2AF2^{RRM2} and FUBP1$^{N\text{-}box}$ was introduced for the subsequent structure determination (Table S6). Overall structural integrity of the chimeric construct and recapitulation of the interaction was confirmed by comparing ^1H–^{15}N HSQC spectra of the chimeric construct to that of the intermolecular complex U2AF2-RRM2–FUBP1$^{N\text{-}box}$ (Figures S3I and S3J). CYANA3 (3.98.15) was used for automated NOE assignments and initial structure calculations.[95] To overcome partial signal broadenings for the resonances at the interface of the two domains, possibly due to the weaker affinity, additional unambiguous intramolecular distance restraints from ^{13}C-NOESY–HMQC and methyl-NOESY spectra were manually assigned and included in the structure calculation.[125] A minimal number of typical hydrogen bonds, which were confirmed by ^{15}N-edited NOESY and secondary structure propensity, was implemented to assist the initial folding during the structure calculation. Dihedral angle restraints were derived from SSP and ^{13}C secondary chemical shifts using TALOS+, including resonances of Ca, Cb, C, H, and N.[96,126] For water refinement, distance restraints from CYANA3 considering an error of ± 0.5 Å are used. Water refinement[127] of the 20 lowest-energy structures (500 initial structures) was performed with ARIA2.3[97] and CNS.[128] The quality of the 10 final structures was evaluated by ProcheckNMR[98] and PSVS.[99] Ensemble structure root mean square (r.m.s.) deviations were calculated using MolMol[100] and the ribbon representations were prepared in PyMOL (The PyMOL Molecular Graphics System, version 1.8.6.0, Schrödinger, LLC). Structural statistics are shown in Table 1.

Scaffold-independent analysis

For the initial screening, the 16 DNA pools of 5-mer DNA (Table S5, #63, IDT), instead of RNA due to their similarity in binding, were generated by introducing a specific nucleotide at a designated position while randomizing the other four positions. Titrations of 100 μM FUBP1 KH domain samples with the different DNA pools (0.5, 1.0, 2.0, and 4.0 molar equivalents of titrant to analyte) were performed at 25°C in NMR buffer (20 mM sodium phosphate, pH 6.5, 50 mM NaCl, 2 mM DTT) containing 10% (v/v) D_2O by recording SOFAST HMQC spectra on a 600 MHz Bruker Avance NMR spectrometer (cryogenic triple-resonance gradient probe). For the comparison and identification of position-specific nucleotide preference, we focused on a subset of 12 representative peaks, which show visibly clear changes in chemical shift (fast-exchange regime) and are therefore involved in binding, for further analysis. CSPs of these peaks were calculated (see above) and the average CSPs of all peaks for each pool were normalized against the largest CSP calculated in the four pools to obtain a score for nucleotide preference at a specific position. The final optimized motifs were verified by comparing the chemical shift changes upon adding either DNA or RNA for all KH domains (Table S5, #67–72).[129]

In vitro binding assays

In vitro *transcription*

All RNA samples were *in vitro* transcribed using T7 RNA polymerase, precipitated by ethanol and purified by denaturing PAGE (12% polyacrylamide gel containing 8 M urea). The DNA templates for *in vitro* transcription are shown in Table S5 (Oligos #59–62). The gel slices were electro-eluted at 250 V in 0.5× TBE. To promote proper folding, the RNA samples were heated to 95°C for 2 min and subsequently snap-cooled on ice before use.

Fluorescent EMSA

In vitro-transcribed RNA was fluorescently labeled by ligation of pCp-Cy5 to the 3' end of the RNA with T4 RNA ligase 2. Subsequently, the reaction was purified using a spin column kit (Norgen Biotek Corp.). For binding studies, 100 nM labeled RNA in 20 mM sodium phosphate, pH 6.5, 50 mM NaCl and glycerol (15% final concentration) was incubated with increasing concentrations of $FUBP1^{N\text{-}box+KH1\text{-}4}$ (amino acids 1–457) for 15 min. Mixtures were loaded onto a 0.7% agarose gel. Gel electrophoresis was performed in 1× TBE buffer at 40 V for 4 h. Detection was performed using a Typhoon 9200 (GE Healthcare Life Sciences) at 649 nm. Data analysis was performed in Image J 2.1.0.[102] Experiments were repeated to estimate the standard deviation of the mean.

Isothermal titration calorimetry

ITC experiments were performed on a MicroCalPEAQ-ITC instrument (Malvern Panalytical) using non-isotopically labeled proteins as analyte sample and titrant or non-isotopically labeled protein as analyte and DNA oligonucleotides as titrant in NMR buffer at 25°C. U2AF2 constructs (concentration 15–30 μM) were titrated with FUBP1 N-terminal constructs (concentration 1.5–3.0 mM); FUBP1 double-KH domain constructs (concentration 20–30 μM) were titrated with DNA oligonucleotides (concentration 200–350 μM, Table S5, #64–66); *in vitro*-transcribed ssRNA (*VPS13D*, 15 μM) was titrated with $FUBP1^{KH}$ (150 μM). Binding affinity analysis was performed using MicroCalPEAQ-ITC Analysis Software (Malvern Panalytical). The standard deviations of the K_D values were estimated based on the differences in triplicate measurements.

BRET

BRET plasmid construction

The donor and acceptor vectors pcDNA3.1-cmyc-NL-GW (Addgene plasmid ID #113446), pcDNA3.1-GW-NL-cmyc (Addgene plasmid ID #113447), pcDNA3.1-GW-mCit, pcDNA3.1-mCit-GW, as well as controls pcDNA3.1-NL-cmyc (Addgene plasmid ID #113442), pcDNA3.1-PA-mCit (Addgene plasmid ID #113443), and pcDNA3.1-PA-mCit-NL-cmyc (Addgene plasmid ID #113444) were kindly provided by the Wanker group (Max-Delbrück-Centrum für Molekulare Medizin, Germany). The GATEWAY entry vectors pDON221 and pDON223 were provided by the Vidal group (Dana Farber Cancer Institute, Boston, MA). All vectors were amplified and full-length sequenced using the primers given in Table S5. Full-length wild-type ORFs being cloned into GATEWAY entry vectors were amplified from a human ORFeome collection.[130] The ORFs were full-length sequenced using primers shown in Table S5. ORFs of *FUBP1*, *SNRNP70*, and *TCERG1* (Table S6) were PCR-amplified with primers #9–10, #27–28, and #33–34, respectively (Table S5) and shuttled into pDON223 using a BP clonase II mix kit (Invitrogen). The Q5 site-directed mutagenesis kit (Invitrogen) was used to produce the following mutants: pDON223-FUBP1_A38D, pDON223-FUPB1_W586R_W615R, and pDON223-FUBP1_1-530aa (Table S6). For BRET experiments, all cDNAs were shuttled from the entry vectors into the BRET destination vectors using LR clonase technology (Invitrogen) according to the manufacturer's protocol. After the LR cloning step, the inserts were partially sequence-confirmed. All primers used are given in Table S5 and all the constructs are listed in Table S6.

Transfection

The human embryonic kidney 293 cells were transfected using Lipofectamine 2000 (Invitrogen) transfection reagent in Opti-MEM medium (Thermo Fisher) using the reverse transfection method according to the manufacturer's instructions. For BRET transfections, cells were seeded at a density of 4.0×10^4 cells per well on a white 96-well microtiter plate (Greiner) in phenol-red-free, high-glucose DMEM media (Thermo Fisher) supplemented with 5% FBS (Thermo Fisher). Transfections were performed with a total amount of 200 ng of DNA per well. If the amount of expression plasmid was less than 200 ng in a well, pcDNA3.1 (+) was used as a carrier DNA to achieve the total of 200 ng.

Experiments

Cells were transfected with plasmids encoding the acceptor (50 ng DNA) and donor (1 ng DNA). The plate was incubated for 2 days at 37°C, 5% CO_2, and 85% relative humidity prior to measurement. All measurements were performed on an Infinite M200 Pro microplate reader (Tecan). First, 100 μl of the medium was aspirated from each well. The mCitrine fluorescence was measured in intact cells (excitation/emission 513/548 nm). Then, coelenterazine h (PJK Biotech GmbH) was added at a final concentration of 5 μM. The cells were briefly shaken and incubated for 15 min inside the plate reader. After incubation, total luminescence was measured first followed by short-wavelength and long-wavelength luminescence measurements using BLUE1 (370–480 nm) and GREEN1 (520–570 nm) filters at 1,000 ms integration time. Corrected BRET (cBRET) ratios were calculated as previously described.[58] In brief, for every transfected protein pair NL-A and mCit-B, the following two control pairs were measured: NL-Stop with mCit-B and NL-A with mCit-Stop. The maximal BRET from both control pairs was subtracted from the actual test pair to correct for donor bleed-through, nonspecific binding to the tags, and background signal.

Saturation assay

For donor saturation experiments 1 ng of donor DNA encoding NL-fused proteins was co-transfected with increasing amounts of acceptor DNA encoding mCitrine-fused proteins (10, 25, 50, 100, 200, 400 ng). Fluorescence, total luminescence, and BRET were measured as described before. BRET measurements were corrected for bleed-through using NL-Stop transfections. Fluorescence and total luminescence measurements were used to estimate the amount of expressed proteins and used to plot acceptor/donor ratios on the *x*-axis.

QUANTIFICATION AND STATISTICAL ANALYSIS

Preprocessing of RNA-seq data

Prior to genomic mapping, remaining adapter sequences were trimmed in RNA-seq data from *FUBP1* KO, *FUBP1-Nbox*^{mut}, and WT control RPE1 cells using Cutadapt v2.4.[108] A minimal overlap of 1 nt between reads and adapter was required and only reads with a length of at least 50 nt after trimming were retained for further analysis (parameters: -O 1 -m 50). Reads were mapped using STAR v2.6.1b,[107] allowing up to 4% of the mapped bases to be mismatched (--outFilterMismatchNoverLmax 0.04 --outFilterMismatchNmax 999) and a splice junction overhang (--sjdbOverhang) of 83 nt for HeLa WT samples and of 158 nt for *FUBP1* KO, *FUBP1-Nbox*^{mut}, and WT control RPE1 cells. Genome assembly and annotation of GENCODE[131] release 31 were used during mapping. Subsequently, secondary hits were removed using Samtools v1.9.[109] Exonic read counts per gene were extracted using featureCounts from the Subread tool suite v1.6.2[110] with non-default parameters --donotsort -s2.

Preprocessing of *in vivo* iCLIP data

Basic quality controls were conducted in FastQC v0.11.8 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc) and reads were filtered based on sequencing qualities (Phred score) in the sample barcode and UMI regions using the FASTX-Toolkit v0.0.14 (http://hannonlab.cshl.edu/fastx_toolkit/) and seqtk v1.3 (https://github.com/lh3/seqtk/). All reads with a Phred score below 10 in the sample barcode or UMI regions were discarded. Reads were de-multiplexed based on the sample barcode, which is found on positions 6–11 of the reads (for 6-nt sample barcodes) or on positions 4–7 (for a 4-nt sample barcode), using Flexbar v3.4.0.[111] Subsequently, barcode regions and adapter sequences were trimmed from read ends using Flexbar, requiring a minimal overlap of 1 nt of read and adapter and adding UMIs to the read identifiers. Reads shorter than 15 nt were discarded. All empty space and slash characters were removed from read identifiers in FASTQ files to prevent all information following them being lost during mapping. The downstream analysis was done as described in Chapters 3.4, 4.1, and 4.2 of Ref. [132]. Genome assembly and annotation of GENCODE[131] release 31 were used during mapping with STAR v2.6.1b.[107] The number of crosslinking events and peaks is given in Table S1. To assess the genomic distribution of iCLIP crosslink nucleotides, we used the following hierarchy: ncRNA > 3′ UTR > 5′ UTR > coding sequence (CDS) > 3′ splice site > 5′ splice site > intron > intergenic (Figure 1B). 3′ and 5′ splice site regions refer to 100 nt upstream/downstream. All other "deep-intronic" regions are called intronic regions.

Metaprofiles for *in vivo* iCLIP data

Four RNA-seq replicates from HeLa cells (imb_koenig_2018_18) served as the source for the identification of spliced introns. Mapping to the genome was performed in STAR v2.6.1b[107] (Table S1). Coordinates and number of unique supporting junction reads ("ureads") of spliced introns were extracted from the SJ file output by STAR containing high-confidence splice junctions. In the following, introns from the SJ file are called "SJ introns". SJ introns had to meet a reproducibility criterion (at least 3 out of 4 replicates). In addition, all overlapping SJ introns were removed. Finally, introns were overlaid with GENCODE release 31 annotation and filtered for level < 3, transcript support level < 4, and gene_type and transcript_type equal to "protein coding". This resulted in 88,375 SJ introns. Branch point (BP) prediction was taken from LaBranchoR.[133] LaBranchoR is based on hg19, liftOver to hg38 was done with the liftOver tool by UCSC.[134] The median distance of BP to 3′ splice sites was 25 nt. 88,008 out of 88,375 SJ introns had an annotated BP. Introns were further filtered for a minimum length of 100 nt and a maximum length of 17,000 nt. Metaprofiles were aligned at the BP. *In vivo* iCLIP replicates for each RBP were summed up and a signal threshold of 10 in the metaprofile region (−200 nt to +50 nt with respect to the BP) was imposed. Crosslinking signals per intron were normalized by "ureads" and averaged per nucleotide over all introns. For display, the normalized signal was smoothed with a Gaussian window function and window size

10. Binding enrichment for RNA maps stratified by intron length and splice site features was calculated by taking the $\log_2$ fold change of the ratio of the area under the curve (AUC) of each feature bin to the AUC in the shortest intron class or class with the weakest splice site feature. The following regions were used for AUC quantification, always with respect to BP: [−100, −25] for FUBP1, [+5, +25] for U2AF2, [−10, +10] for SF1, and [−30, −10] for SF3B1. The minimum signal in each region served as a background proxy and was taken as the lower horizontal boundary in which the AUC was calculated. For RNA maps stratified by GC content, the average GC content in the exon was contrasted to the average GC content in the first 100 nt of the downstream intron. Signal values for RNA maps aligned at 5′ splice sites were not smoothed but normalized by the average signal in the first 100 nt of the intron. RNA maps conditioned on exon rank: annotation of exons and downstream introns was extracted from GENCODE release 31. BPs were annotated as described above. SJ introns were matched to introns. Duplicated matches were resolved such that the intron with the shortest upstream exon was taken. Five exon rank classes were extracted: 1st exon, exon ranks in [2,5), [5,12), [12, 144] and second to last exon. In comparison to all other RNA maps, crosslinking signals per intron were normalized to the total crosslinking signal in the last 100 nt upstream of the 3′ splice site. "ureads" correlates with exon rank and was thus not suitable as a normalization factor. RNA maps conditioned on exon GC content: upstream exons were identified as for exon rank RNA maps. Total exon GC content over exon length was extracted. Bins are as follows: [0.07,0.41], (0.41, 0.46], (0.46, 0.53], (0.53, 0.6], (0.6, 0.91]. RNA maps condition on intron GC content: total intron GC content over intron length was extracted. Bins were as follows: [0.14, 0.36], (0.36, 0.4], (0.4, 0.46], (0.46, 0.55], (0.55, 0.9]. RNA maps for fixed intron length/differential GC content architecture followed by subsequent conditioning on differential GC content/intron length. Here, RNA binding profiles were first stratified on one class of intron length/differential GC content architecture, followed by stratification on all levels of the other factor. Binding for all RNA maps was quantified based on AUC as described above. Analyses were performed in R v4.1.1.[104]

iCLIP binding site definition (peak calling)

Binding site definition for *in vivo* iCLIP was done with PureCLIP v1.3.1. on merged replicates.[112] PureCLIP was issued with the options -iv 'chr1;chr2;chr3;' -ld -nt 4. The crosslink sites identified by PureCLIP were post-processed as previously described.[132] In detail, individual crosslink sites within a distance of 5 nt were clustered into binding regions. The binding regions were resized to obtain binding sites of a uniform width. To compare binding sites of different RBPs, we opted for 5-nt binding sites (i.e., 2 nt on either side of the position with the maximum signal) for all of the RBPs investigated (FUBP1, U2AF2, SF3B1, SF1, PTBP1). Isolated crosslink sites and binding regions of 2 nt were removed. Binding regions $\leq$ 5 nt were centered on the position with the maximum crosslink signal and extended by 2 nt on either side. Binding regions > 5 nt were divided into regions of 5 nt, by iteratively screening for the maximum signal and extending of 2 nt on either side, excluding an overlap between binding regions. Finally, at least three positions with crosslink events were required to only keep binding sites with sufficient support. To ensure sufficient support of binding sites in the individual replicates of the experiment, a reproducibility filter was applied. In order to consider the varying number and size of replicates for each experiment, we filtered for those binding sites with a total number of crosslink events higher than the 10% percentile of the distribution of crosslink counts in the single replicate. In addition, a minimum of two crosslink events was required if the 10% percentile in the replicate was below this threshold. This was required in at least two out of three, three out of four and three out of five replicates depending on the number of replicates available for the respective experiment. The numbers of called binding sites per protein are given in Table S1.

Saturation analysis

Spliced introns were identified from four RNA-seq replicates in HeLa cells (imb_koenig_2018_18) as described above. Introns were retained if they were longer than 200 nt, and if the 5′ splice site windows (the last 50 nt of the exon plus the first 75 nt of the intron) and 3′ splice site windows (the last 200 nt of the intron plus the first 20 nt of the exon) were not overlapping. 3′ splice sites overlapping to noncoding and long noncoding RNAs were excluded, resulting in 98,328 3′ splice sites. These splice sites were binned into percentiles, based on "ureads" (splice site usage) averaged over replicates. RBP binding sites were assigned to curated 3′ splice sites (the last 200 nt of the intron), requiring full overlap. For each bin, the percentage of 3′ splice sites with at least one binding site for the specific RBP was calculated (Figure 1E).

Motif enrichment for *in vivo* iCLIP

Introns were defined based on GENCODE annotation (release 31). Annotation was filtered for level < 3, transcript support level < 4, and gene_type and transcript_type equal to "protein coding", resulting in 202,623 introns. BP annotation was done as specified above. 200,199 out of 202,623 had an annotated BP. Introns were further filtered for overlaps and for having a length of at least 250 nt upstream of the defined BP. The length requirement was set to ensure that the main position of FUBP1 binding was not confounded with the 5′ splice site signal. FUBP1 binding sites (n = 854,404) were filtered for positioning within a 150-nt window upstream of the BP, resulting in 167,408 binding sites. Binding sites were ranked by their normalized signal, that is, the signal in the extended binding site (5 nt ± 5 nt) over total intron signal over intron length. Disjunct 4-mer frequencies were counted in the top/bottom 20% binding sites based on normalized signal to account for overall crosslinking preferences. Additionally, non-bound intronic regions in introns hosting the top 20% FUBP1 binding sites were also considered as an alternative background set. Here, disjunct 4-mer frequencies were calculated for all non-bound intronic regions, excluding a 20-nt region downstream of the 5′ splice site and a 150-nt region upstream of the BP. Enrichment was defined as the distance from each data point to the diagonal in a scatterplot

comparing the top 20% versus bottom 20% binding sites and, alternatively, non-bound intronic sequences. Analyses were performed in R v4.1.1.[104]

Motif enrichment upstream of branch points

Introns were extracted and BP annotated as above (200,199 introns left). Introns were further filtered for a minimum length of 500 nt and disjunct 200-nt windows upstream of the BP, resulting in 151,836 introns. Disjunct 4-mer frequencies were calculated in a position-wise manner in a 200-nt window upstream of the BP. Average background motif frequencies were calculated in a 100-nt long window 100 nt downstream of the 5′ splice site. Enrichment was defined as the distance from each data point to the diagonal in the scatterplot of position-wise frequencies versus average background frequencies.

Abundance of FUBP1 motif at 3′ splice sites

Disjunct motif occurrences were counted in a 75-nt long window 25 nt upstream of the BP. The background distribution was derived as the occurrences of nine randomly drawn motifs of length 4, repeated 100 times.

Analysis of *in vitro* iCLIP data

All samples were merged for binding site definition (peak calling) across replicates and conditions. Each *in vitro* transcript was divided into 9-nt windows, always shifted by one nucleotide. Windows were sorted by total signal and, while excluding overlapping peaks, generating a candidate. A negative binomial distribution was fit (maximum likelihood fit) to the signals on the candidate peak list. All peaks with a total signal exceeding the 90% quantile of the theoretical distribution were retained for final processing (109 peaks, see GEO record GSE220183). The background ranges were the *in vitro* transcript regions minus extended peaks (9 nt ± 5 nt). For quantifying the binding differences between conditions, replicates were averaged. Peak signals were normalized against background signals. RNA maps were based on 21 3′ splice sites present in the *in vitro* transcripts. To correct for differences in expression, nucleotide-wise signals were normalized by total *in vitro* transcript signals. Subsequently, signals were summarized per nucleotide by the 75% quantile. Replicates were averaged and subjected to Gaussian window smoothing with window size 10 before display. All analyses were performed in R v4.1.1.[104]

Analysis of oligo in vitro *iCLIP*

All data was normalized according to the total signal of all available spike-ins. Values were then extracted either per nucleotide or by binding site. Binding site positions were taken from overlays with *in vivo* U2AF2 binding sites in the intronic part of the oligonucleotide. 1,831 oligonucleotides harbored an U2AF2 binding site in the intronic part (see GEO record GSE220183). If multiple binding sites were present, that with the highest average signal in the U2AF2 samples was taken as representative. For quantifying the addition of FUBP1 on U2AF2 binding sites, only those binding sites with signal greater than the 25% quantile in one of the three replicates were considered, resulting in 1,504 binding sites. The absolute number of disjunct occurrences of the FUBP1 motif set ("TTTT" and all combinations of "TTT" and either one "A" or one "G") was counted in a 75-nt long region located 25 nt upstream of the BP. All analyses were performed in R v4.1.1.[104]

Intron length analyses of RNA-seq data

Splicing changes of *FUBP1* KO and *FUBP1-Nbox*^{mut} were analyzed with MAJIQ v2.2[135,136] with default parameter settings. MAJIQ outputs local splice variations (LSV), which were filtered as follows: for each LSV, the top two junctions in terms of absolute difference in junction usage (delta percent selected index, |ΔPSI|) were taken as representative LSVs. At least one of these two junctions needed to have an absolute ΔPSI > 0.1 and a detection probability > 0.9 (skipped for control events). Subsequently, events were filtered for exon-skipping events. Each cassette exon was then annotated with the upstream and downstream intron: genomic coordinates of the upstream/downstream intron were immediately defined in "source"/"target" events. The genomic coordinates of the respective other intron were extracted from annotation (GENCODE release 31). Overlapping cassette exons were resolved such that the event with the largest |ΔPSI| was retained (Table S3). A two-tailed Wilcoxon rank-sum test was used to assess statistical significance.

ENCODE data analysis

We retrieved raw RNA-seq data derived from an shRNA-knockdown experiment for FUBP1 in the cell line K562 from the ENCODE data portal (https://www.encodeproject.org/), using accession numbers ENCSR608IXR (*FUBP1* KD) and ENCSR260BQC (control). Alignment was performed in STAR (version 2.7.8a)[107] with standard ENCODE options. We applied MAJIQ v2.3[135,136] to identify and quantify cassette exons in the RNA-seq data. First, a splice graph was built on the BAM files and the GENCODE gene annotation (v38, human genome version hg38). Then, the difference in junction usage between knockdown and control samples was calculated (as ΔPSI). Next, alternative splicing events such as cassette exons (CEs) were categorized and quantified in the splicing graph using MAJIQ Modulizer. Probabilities were calculated for each junction, testing for |ΔPSI| > 0.05 (probability changing [P_s]) and |ΔPSI| < 0.02 (probability non-changing [P_n]). The MAJIQ Modulizer output was then processed in R, filtering for significantly regulated CEs and a control group with unregulated CEs. A CE is defined as significantly regulated if $|\Delta PSI| \geq 0.055$ for all junctions, $P_s \geq 0.9$ for at least one junction pair (inclusion junction + skipping junction), the sign within both junction pairs is inverse, and within the junction pairs the lower |ΔPSI| is at least 50% of the higher |ΔPSI|. A CE is considered to be unregulated if $P_n \geq 0.5$ and $|\Delta PSI| \leq 0.02$ for all junctions. Overall, this resulted in a total of 173 significantly regulated CEs and a control group with 1,910 unregulated CEs for further

analysis. To categorize CEs into more included and less included, a representative ΔPSI was chosen for each CE based on the maximum |ΔPSI| of both inclusion junctions. Based on this, there were 30 more-included and 143 less-included exons.

Splicing changes upon FUBP1 LoF mutations

Significant differentially spliced exon-skipping events upon (i) loss-of-function (LoF) mutations of *FUBP1* in low-grade gliomas (37 events), (ii) in *FUBP1* siRNA knockdown in U87MG cells (109 events) and (iii) LoF mutations of other splicing factors (433) were extracted from Seiler et al.[1] Junction lengths comprise the upstream intron, the skipped exon and the downstream intron. A two-tailed Wilcoxon rank sum test was used to assess statistical significance.

Mutations in FUBP1 in cancer patients

We searched multiple databases to identify disease-related mutations within the *FUBP1* gene. We focused on the minimal binding interface to U2AF2 (FUBP1 amino acids 25–56) to find mutations that potentially abolish the interaction with the U2AF2 RRM2 domain. The following databases were used: ICGC Data Portal,[137] cBioPortal,[138,139] Exac,[140] Cosmic,[141] GDC Data Portal,[142] gnomAD,[140] and ClinVar.[143] All cancer-related mutations in *FUBP1* in the observed region and the underlying cancer type are listed in Figure S4B.

Scoring of splice site features

3′ and 5′ splice site strength was scored with MaxEnt scan.[144] Py tract strength was determined as follows: a 39-nt region upstream of the AG dinucleotide at the 3' splice site was screened with sliding windows of increasing length (width 5–30 nt) to identify the window with the highest Py tract strength. The Py tract strength of each window was calculated as the X^2 test statistic with 1 degree of freedom, comparing the observed number of pyrimidines with the expected number based on the assumption of a uniform nucleotide distribution. In addition, candidate Py tracts were required to end within 10 nt upstream of the AG dinucleotide. Using this approach, the median length of identified Py tracts was 16 nt. BP strength was assessed according to the U2 binding energy, that is, the number of hydrogen bonds between the candidate sequences and the BP binding sequences in the U2 snRNA. Hydrogen bonds form between A:T (2 bonds), G:C (2 bonds), and G:U (1 bond; in fact also 2 bonds, but punished for being a wobble base pair) with the BP nucleotide bulging out and being omitted from the pairings. The Vienna RNA package v2.4.17[113] (RNAduplex) was used to determine the optimal hybridization structure between U2 snRNA sequences (GUGUAGUA) and the motif (position −5 to +3, excluding the BP nucleotide). Predicted binding energy was the determined sum of hydrogen bonds forming between complementary motifs and U2 snRNA nucleotides.

Evolutionary analyses

We annotated the domain architecture of FUBP1 using the function annoFAS provided in the FAS package[106] (https://github.com/BIONF/FAS). The domain architecture-aware phylogenetic profile of FUBP1 across 174 mammals, 274 non-mammalian vertebrates, 277 invertebrates, 410 fungal species, 94 protozoa, and 145 plants was generated with the targeted ortholog search tool fDOG (https://github.com/BIONF/fDOG)[145] using the human FUBP1 (UniProt: Q96AE4) as a seed. fDOG was run with the options --minDist class, --maxDist phylum, –checkCoorthologsRef, and --countercheck. *Homo sapiens* (GenBank: GCF000001405) served as the reference taxon. Intron length and GC content information was extracted based on the respective gff and fasta files downloaded from NCBI RefSeq Genome. Intron length estimates and motif searches were performed in R v4.0.5. A/B box presence in the human proteome was determined as follows: in brief, we used the shell command *grep* to search for the regular expression "[ST][AK][QA]W..YY[RK]" in 19,519 human proteins encoded in the NCBI RefSeq Genome assembly GCF_000001405.39. The resulting three hits were NCBI: XP_011540693.1 (FUBP1, 2 motif instances), NCBI: NP_003925.1 (FUBP3, 1 motif instance), and NCBI: NP_001353228.1 (KHSRP, 3 motif instances). For counting FUBP1 motif occurrences across species, intron definitions were extracted for all the species investigated and motifs were counted in a 25-nt window located 25 nt upstream of the 3′ splice site.

Analysis of RBP crosslinking to snRNAs

In vivo iCLIP data from FUBP1, U2AF, SF1, SF3B1, and PTB was remapped to a custom database consisting of snRNAs, tRNAs, and rRNAs using STAR v2.7.3a.[107] Specifically, RNU1-1, RNU2-1, RNU4-1, RNU6-1, RNU5D-1, RNU7-1, RNU11, RNU12, RNU4ATAC, and RNU6ATAC were included. tRNA coordinates were retrieved from GtRNAdb (data release 19). "hg38-tRNAs.fasta", containing 429 high-confidence tRNA annotations, was downloaded. Because tRNAs are quite similar when stratified on their carried amino acid, one representative tRNA was selected per amino acid (tRNA with "1-1" in the name). In summary, this resulted in 22 tRNAs. Finally, the following rRNAs were added: 12S_gi, 16S_gi, 18S_gi, 28S_gi, 5.8S_gi, and 5S_gi. Mapping steps were performed as follows: all sequences were furnished with one additional base upstream of the sequence with the rationale of being able to display iCLIP coverage of reads starting directly at the 5' end of the sequence. tRNAs and snRNAs were furnished with the actual base upstream of the sequence. rRNAs were furnished with an "N". Reads were mapped per replicate with STAR v2.7.3a using the settings described above for *in vivo* iCLIP samples. Few reads were mapped to the minus strand and thus removed. Uniquely mapping reads were subjected to duplicate removal based on identical UMIs (--method unique) using UMI-tools v1.0.0.[146] Based on the remaining reads, iCLIP coverage profiles were exported as well as count tables containing the number of reads overlapping the genomic ranges of the defined RNAs.

Subnuclear distribution of FUBP1-bound genes

The subnuclear spatial distribution for introns in HeLa cells was taken from Tammer et al.[64], in which Chrom3D, a 3D genome-modeling tool that integrates 3DHi-C data and ChIP-seq data was used to assign distances from the nuclear center for topologically associated domains. The distance from the nuclear center is described by five concentric radial scopes where 1-to-5 point to the center–periphery axis. Our *in vivo* iCLIP data from SF3B1, FUBP1, and U2AF2 was then overlaid with the reported introns and the percentage of bound introns was counted. Enrichment was calculated as the percentage of bound introns in each radial scope compared to the first.

Mathematical modeling

Topology of the exon definition model

Splicing reactions are catalyzed by the spliceosome, which recognizes splice site sequences and forms a catalytically active higher-order complex across introns. To model this process, we considered that human spliceosomes frequently operate by a so-called "exon definition" mechanism, in which the pioneering spliceosome subunits U1 and U2 cooperatively bind to splice sites flanking an exon before the final cross-intron complex is formed during spliceosome maturation.[86] Because the initial binding of U1 and U2 plays a decisive role in splicing decisions,[86] we model only the initial exon definition step and assume the corresponding binding patterns determine splicing outcomes, as described below.

In the model pre-mRNA, none of the three exons are bound ("defined") by the spliceosome (white boxes), therefore this state is denoted "P0_0_0" (Figure S7F) with the notation "_" indicating the presence of an intron. In the model, the pre-mRNA (P0_0_0) is synthesized at a constant rate s. The spliceosome can bind reversibly to each of the exons with on-rates k_1, k_2, and k_3. For instance, from P0_0_0 we can obtain P1_0_0, P0_1_0, and P0_0_1 through binding to the first, second, and third exon, respectively. Subsequent binding is possible; for example, P1_0_1 can be generated from P1_0_0 with the rate constant k_3. In total, there are eight spliceosomal binding states, including the fully bound state (P1_1_1), in which all exons are defined. All binding reactions are assumed to be reversible, i.e., k_4, k_5, and k_6 are the dissociation rate constants and the reverse of k_1, k_2, and k_3, respectively. For example, in state P1_1_0, spliceosome dissociation from exon 1 with the rate constant k_4 yields the species P0_1_0.

Depending on the exon definition states, splicing decisions are made, and irreversible splicing reactions are possible. For a splicing event to occur, we consider that both exons flanking a future splice junction must be defined. For instance, skipping of exon 2 is possible from P1_0_1 and occurs with the rate constant i_{12}. Likewise, splicing of the first intron occurs from the species P1_1_0 and P1_1_1 (rate constant i_1), and splicing of the second intron from P0_1_1 and P1_1_1 (rate constant i_2). The inclusion isoform is generated in two steps, that is, from the subsequent removal of introns 1 and 2 in random order: from the binding state P1_1_1, intron splicing generates two alternative intermediates in which either of the introns is already spliced (P1_11 or P11_1) and the retained intron can be further spliced in a subsequent reaction. Splicing of the partially defined species P1_1_0 and P0_1_1 yields the species P11_0 and P0_11; in these, the spliceosome can further reversibly bind exons 3 and 1, respectively, and undergo a second splicing reaction toward inclusion. In the model, all terminal splice products are subject to degradation (k_{incl}, degradation rate constant of inclusion; k_{skip}, skipping; k_{dr1}, first intron retention; k_{dr2}, second intron retention). The degradation rate constant of the full intron retention isoform is the sum of k_{dr1} and k_{dr2}, reflecting that either intron may contain a destabilizing premature stop codon. Model species that can be bound or spliced further (P0_0_0, P1_0_0, P0_1_0, P0_0_1, P1_1_0, P1_0_1, P0_1_1, P1_1_1, P0_11, P1_11, P11_0, P11_1) are not subject to degradation, but they can be exported from the nucleus with the rate constant k_{ret}. This reaction reflects that there is a limited time window for splicing to occur, the intermediates otherwise being terminally frozen in the corresponding intron retention state. The ordinary differential equations of the model are given in Table S4.

Topology of the intron definition model

Because a subset of human genes are spliced by an intron definition mechanism, we also considered this scenario in a modified version of our splicing model. In contrast to the exon definition model, the 5′ and 3′ splice sites of an exon can be bound independently of one another in the intron definition model. Furthermore, splicing of an intron is possible as soon as both splice sites flanking this intron are defined. Hence, definition of two splice sites is sufficient for splicing to occur, whereas in the exon definition model four splice sites need to be defined (3′ and 5′ splice sites of the two flanking exons). For the intron definition model, we use a notation for binding state similar to that for exon definition. For instance, for consistency, we assigned the state in which no spliceosome component is bound as P0_0_0. For spliceosome binding to exons 1 and 3, we again considered a single binding reaction, as only the splice sites flanking the intron of interest are relevant for splicing. Hence, a transition from "0" to "1" in the first position (e.g., P0_0_0 to P1_0_0) represents a spliceosome binding state downstream of exon 1 (5' of the first intron), and "0" to "1" in the third position indicates binding upstream of exon 3 (3' of the second intron). For exon 2, we treat splice-site binding as two separate events. We use "0" to denote no binding, "a" for upstream binding (e.g., P0_a_0), "b" for downstream binding (e.g., P0_b_0), and "1" for both U2 and U1 being simultaneously bound (e.g., P0_1_0). Again, the presence or absence of "_" indicates whether or not the intron is removed. We adopted the same parameter notation, that is, k_1/k_4 and k_3/k_6 to describe binding/dissociation at exons 1 and 3, respectively. The new parameters k_{2a}/k_{5a} (upstream) and k_{2b}/k_{5b} (downstream) were introduced to represent spliceosome binding/dissociation around exon 2. There are a total of 16 spliceosomal binding states in the intron definition model, with the following additional states not part of the exon definition model: P0_a_0, P0_b_0, P1_a_0, P1_b_0, P0_a_1, P0_b_1, P1_a_1, and P1_b_1. If both splice sites flanking a future splice junction are defined, splicing decisions, implemented as irreversible splicing reactions in the model, can occur. Skipping of exon 2 is possible from P1_0_1 and occurs with the rate i_{12}. Splicing of the first intron occurs from species P1_a_0, P1_1_0,

P1_a_1, and P1_1_1 (rate i_1), and splicing of the second intron occurs from P0_b_1, P0_1_1, P1_b_1, and P1_1_1 (rate i_2). The inclusion isoform is generated in two steps: first, intron 1 or 2 is spliced from P1_1_1, generating P1_11 or P11_1, respectively. Second, the retained intron can be further spliced in a subsequent reaction. Splicing of the partially defined species P1_a_0, P1_1_0, P0_b_1, and P0_1_1 yields the species P1a_0, P11_0, P0_b1, and P0_11, respectively. To these, the spliceosome can bind further reversibly with the association rate constants k_1, k_{2a}, k_2, and k_3 (depending on the site of binding), and if the species P1_11 or P11_1 are formed, a second splicing reaction toward inclusion can occur. All terminal splice products are subject to degradation, for which we adopted the same assumptions and notation as for the exon definition model. Again, model species that can be bound or spliced further (P0_0_0, P1_0_0, P0_a_0, P0_b_0, P0_1_0, P0_0_1, P1_1_0, P1_a_0, P1_b_0, P1_0_1, P0_a_1, P0_b_1, P0_1_1, P1_a_1, P1_b_1, P1_1_1) can be exported from the nucleus with the rate constant k_{ret}. The ordinary differential equations of the model are given in Table S4.

Model simulation and analysis

The differential equations were implemented in Matlab 2020b and solved using ode15s. To analyze splicing outcomes, we assumed a steady state, and performed numerical simulations over long time periods (t = 1,000,000 min) to ensure that the concentrations of the model species remained constant. Thus, we consider an RNA sequencing experiment, in which gene expression was measured in a stationary cell population in the absence of any external perturbation. As a measure of splicing outcome, we used the steady-state concentrations of inclusion and skipping (see also below).

Genome-wide splicing modeling by parameter sampling

The exon definition model consists of 15 kinetic parameters which belong to the following classes of reactions: spliceosome binding (k_1, k_{2a}, k_{2b}, k_3), spliceosome dissociation (k_4, k_{5a}, k_{5b}, k_6), splicing catalysis (i_1, i_2, i_{12}), and others, which are rates of pre-mRNA synthesis (s), mRNA degradation (k_{int}, k_{skip}, k_{dr1}, k_{dr2}), and terminal intron retention (k_{ret}). The values of these parameters were unknown and likely greatly differ between exons in the human genome. To mimic the heterogeneity of exons in the human genome and to assess the robustness of our simulation results, we randomly sampled all kinetic parameters in our model 10,000 times. As a reference parameter set, all parameter values were set to 1, except for k_{ret}, k_{incl}, and k_{skip}, which were set to 0.01 to ensure low levels of intron retention that are typically observed in RNA sequencing datasets. We sampled each parameter in the model within a +/-sevenfold range around this reference using Latin hypercube sampling (lhsdesign command in Matlab). We performed simulations for each parameter realization and calculated PSI = inclusion / (inclusion + skipping) as a measure of alternative splicing. We obtained a PSI distribution between 0 and 1 that closely resembled the experimentally measured genome-wide PSI in control cells. The same procedure was applied for intron definition, with the only difference being the number of parameters involved -17 in this case. These kinetic parameters belong to the following classes of reactions: spliceosome binding (k_1, k_{2a}, k_{2b}, k_3) and spliceosome dissociation (k_4, k_{5a}, k_{5b}, k_6); the remainder are identical to those used for exon definition.

Modeling FUBP1 knockout effects

To reproduce the *FUBP1* KO data, we implemented two distinct assumptions about the mechanism of action of FUBP1: that FUBP1 affects late spliceosomal catalysis (i.e., the rate constants i_1, i_2 and/or i_{12}), or that FUBP1 affects early spliceosomal binding (i.e., the rate constants k_1–k_6). For both mechanistic assumptions, we considered that FUBP1 predominantly binds long introns (Figure 6A). When simulating the effect *FUBP1* KO has on splicing catalysis (model 2 in Figure 6A), we assumed that the splicing of short introns is unaffected, but that KO selectively reduces the splicing rate for the excision of long introns 3.5-fold compared to control. To reflect different combinations of long and short introns, we considered three scenarios in the *FUBP1* KO simulations: (i) for the simulation of cassette exons flanked by two long introns, we assumed that the *FUBP1* KO slows all three splicing reactions in the model, that is, the excision of intron 1, excision of intron 2 and exon skipping (i_1, i_2, and i_{12} are changed). (ii) For exons flanked by one short and one long intron, it was assumed that the splicing rate of the short intron is unaffected by *FUBP1* KO, whereas splicing rates of the long intron and skipping are reduced. The long intron was either considered to be located upstream of the alternative exon (ii.a: i_1 and i_{12} are changed) or downstream (ii.b: i_2 and i_{12} are changed). In either case, the skipping reaction was considered as an FUBP1-dependent, long-range splicing event and was therefore perturbed in the *FUBP1* KO simulation (i_{12} is changed). (iii) The third hypothetical scenario, in which an alternative exon is flanked by two short introns, was not explicitly considered in our simulations, as the model would predict no PSI change upon *FUBP1* KO in this case. For each parameter sample (hypothetical exon), the KO scenarios i, ii.a, and ii.b were implemented separately, resulting in three sets of 10,000 KO simulations. For each of these, the PSI changes upon *FUBP1* KO were calculated [ΔPSI = PSI(KO) – PSI(control)], and the corresponding ΔPSI distribution (Figure 6B) agrees well with the experimental observation in RNA sequencing experiments. In the alternative *FUBP1* KO implementation (model 1 in Figure 6A), we assumed that FUBP1 promotes initial U2 binding to the 3′ splice site. Because the 3′ splice site marks the downstream end of an intron, we assume that the *FUBP1* KO reduces spliceosome binding to exons located downstream of long introns. In our model, a long intron 1, therefore, results in a reduced exon 2 definition rate upon *FUBP1* KO (k_2 changed 1.7-fold compared to control). Likewise, a long intron 2 diminishes exon 3 definition (k_3 changed 1.7-fold upon *FUBP1* KO). These perturbations were implemented alone (one long and one short intron), or in combination (two long introns), and the corresponding ΔPSI distributions across all 10,000 parameter realizations are shown in Figure 6B. The perturbation in binding parameters (k_2, k_3) was chosen to be smaller (1.7-fold) than the effect on splicing parameters (3.5-fold, model described above) to adjust for similar-sized effects on splicing in both implementations. In contrast to the *FUBP1* KO RNA sequencing data, these spliceosome binding simulations predict opposite PSI changes for short introns being located upstream or downstream of the alternative exon. Hence, a model in which FUBP1 enhances the catalytic excision of long introns explains the *FUBP1* KO data better when compared to a model in which FUBP1 primarily helps to

recruit the pioneering U2 subunit to the 3′ splice site. The same *FUBP1* KO simulations were also implemented in the intron definition scenario. Here, the effect of FUBP1 on spliceosome binding (model 1 in Figure 6A) was assumed to affect the k_{2a} parameter for a long upstream intron and k_3 for long downstream introns. If both introns are long, FUBP1 influences both k_{2a} and k_3. The effect of FUBP1 on splicing catalysis (model 2 in Figure 6A) in the intron definition model was implemented in the same way as described above for the exon definition model. For FUBP1-based mechanisms of action, that is, binding and catalysis effects, very similar results were observed for the intron and exon definition scenarios (Figure S7G). Hence, the model's prediction that FUBP1 affects splicing catalysis is robust and does not depend on the mechanism of splicing decision making.

2.3.1 Supplementary material

Molecular Cell, Volume *83*

Supplemental information

FUBP1 is a general splicing factor facilitating 3′ splice site recognition and splicing of long introns

Stefanie Ebersberger, Clara Hipp, Miriam M. Mulorz, Andreas Buchbender, Dalmira Hubrich, Hyun-Seo Kang, Santiago Martínez-Lumbreras, Panajot Kristofori, F.X. Reymond Sutandy, Lidia Llacsahuanga Allcca, Jonas Schönfeld, Cem Bakisoglu, Anke Busch, Heike Hänel, Kerstin Tretow, Mareen Welzel, Antonella Di Liddo, Martin M. Möckel, Kathi Zarnack, Ingo Ebersberger, Stefan Legewie, Katja Luck, Michael Sattler, and Julian König

Figure S1

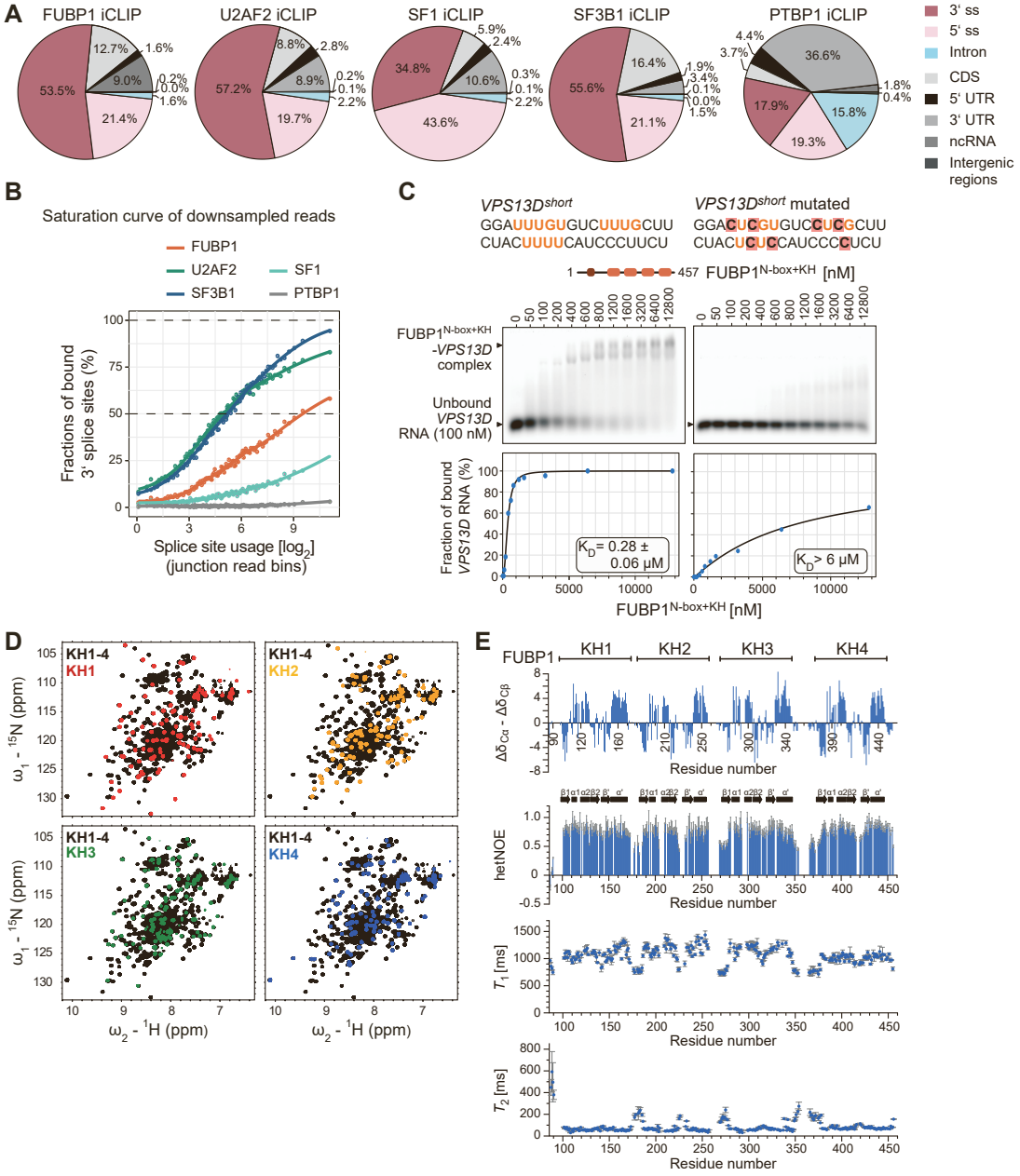


Figure S1. FUBP1 binding at 3' splice sites and RNA binding of KH domains (related to Figure 1B, 1E and 2B-D)

(A) Distribution of binding sites across transcript regions for FUBP1 (n = 854,404), U2AF2 (n = 914,221), SF1 (n = 99,305), SF3B1 (n = 1,694,991), and PTBP1 (n = 127,450) iCLIP in HeLa cells (normalized for total transcript length). 3' and 5' splice site (ss) refer to 100 nt upstream and downstream of exons, respectively. CDS, coding sequence; UTR, untranslated region.

(B) Saturation analysis on downsampled iCLIP data (FUBP1, ~57,000,000 crosslink events; SF3B1, ~68,000,000; U2AF2, 54,000,000; SF, 58,000,000; PTB, 49,000,000), where the iCLIP data for each splicing factor have approximately the same sequencing depth.

(C) Fluorescent electrophoretic mobility shift assay (EMSA) experiment on recombinant FUBP1$^{N\text{-}box+KH}$ (aa 1–457, 50 nM–12.8 μM) binding to a shortened 36-nt RNA fragment from *VPS13D* (*VPS13D*^{short}, 100 nM) (left). Agarose gel image (top) and quantification (bottom) with fitted curve show FUBP1–RNA binding in the nanomolar range (dissociation constant [K_D] = 0.28 ± 0.06 μM). Agarose gel of a fluorescent EMSA experiment on recombinant FUBP1$^{N\text{-}box+KH}$ (aa 1–457, 50 nM–12.8 μM) binding to *VPS13D*^{short} mutated (100 nM) with U-to-C mutations in U-rich stretches affording greatly reduced binding (right).

(D) Overlays of the ^{1}H–^{15}N heteronuclear single quantum coherence (HSQC) spectra of FUBP1 KH1–4 (black) with single KH domains (KH1, red; KH2, yellow; KH3, green; KH4, blue). Nuclear magnetic resonance (NMR) experiments of FUBP1^{KH} (KH1–4) show excellent spectral quality, despite the high molecular weight (~40 kDa), allowing most of the backbone chemical shifts to be assigned (310 out of 371 residues).

(E) $^{13}C_\alpha$ and $^{13}C_\beta$ secondary chemical shifts and ^{15}N relaxation experiments: {^{1}H}-^{15}N heteronuclear nuclear Overhauser effect (NOE), T_1, T_2, of the four KH domains of FUBP1. Folded KH domains exhibit more rigidity (NOE ~ 0.9, T_1 ~ 1s, T_2 ~ 60 ms) whereas linker regions are more flexible (lower NOE, lower T_1, higher T_2).

Figure S2

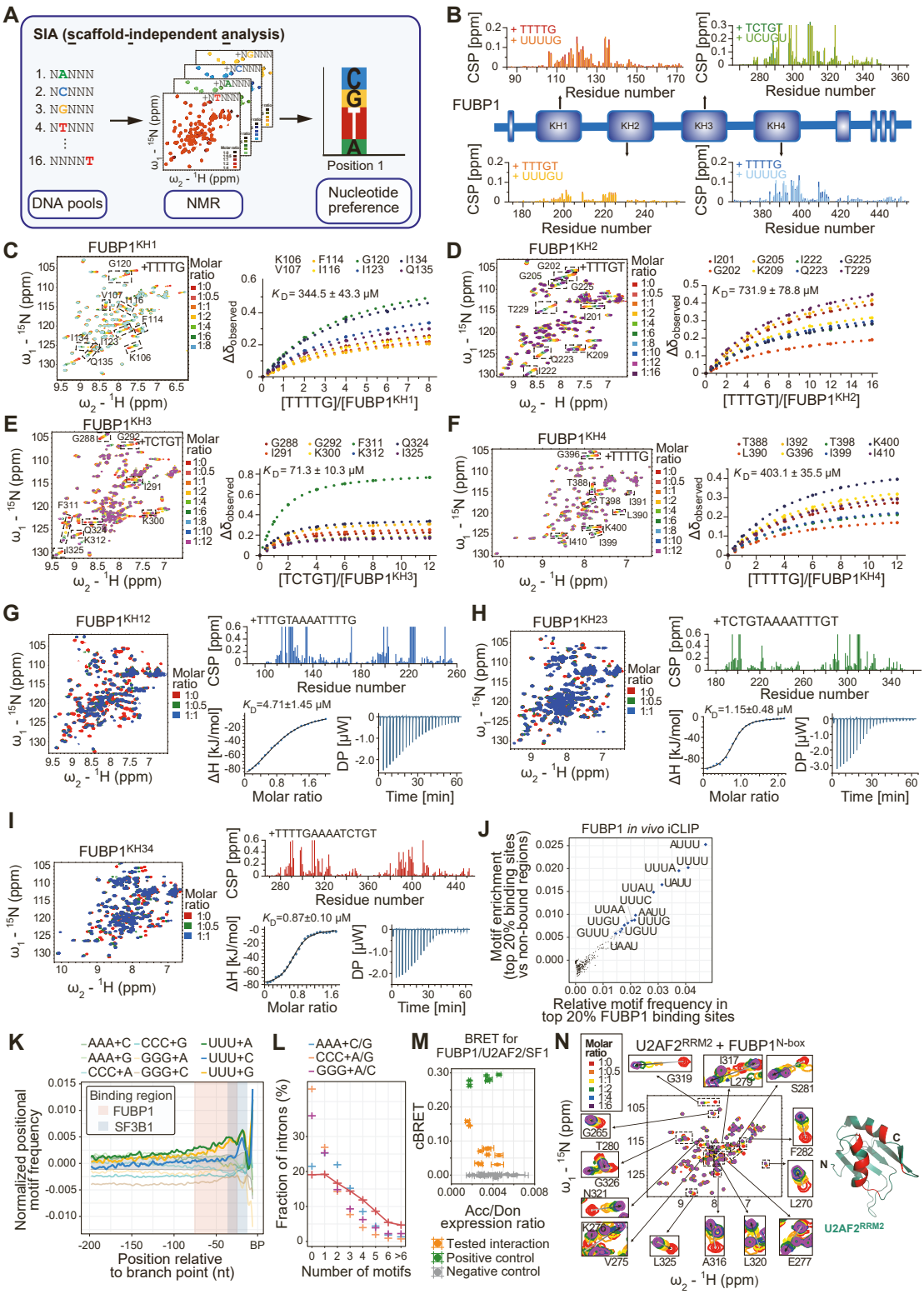

A
SIA (scaffold-independent analysis)
1. NANNN
2. NCNNN
3. NGNNN
4. NTNNN
16. NNNNT
DNA pools
NMR
Position 1
Nucleotide preference
B
+ TTTTG
+ UUUUG
+ TCTGT
+ UCUGU
+ TTTGT
+ UUUGU
FUBP1
KH1
KH2
KH3
KH4
CSP [ppm]
Residue number
C
FUBP1KH1
+TTTTG
Molar ratio
K_D = 344.5 ± 43.3 µM
[TTTTG]/[FUBP1KH1]
D
FUBP1KH2
+TTTGT
K_D = 731.9 ± 78.8 µM
[TTTGT]/[FUBP1KH2]
E
FUBP1KH3
+TCTGT
K_D = 71.3 ± 10.3 µM
[TCTGT]/[FUBP1KH3]
F
FUBP1KH4
+TTTTG
K_D = 403.1 ± 35.5 µM
[TTTTG]/[FUBP1KH4]
G
FUBP1KH12
+TTTGTAAAATTTTG
K_D=4.71±1.45 µM
ΔH [kJ/mol]
DP [µW]
Molar ratio
Time [min]
H
FUBP1KH23
+TCTGTAAAATTTGT
K_D=1.15±0.48 µM
I
FUBP1KH34
+TTTTGAAAATCTGT
K_D=0.87±0.10 µM
J
FUBP1 in vivo iCLIP
Motif enrichment (top 20% binding sites vs non-bound regions)
Relative motif frequency in top 20% FUBP1 binding sites
K
Binding region
FUBP1
SF3B1
Normalized positional motif frequency
Position relative to branch point (nt)
L
Fraction of introns (%)
Number of motifs
M
BRET for FUBP1/U2AF2/SF1
cBRET
Acc/Don expression ratio
Tested interaction
Positive control
Negative control
N
U2AF2RRM2 + FUBP1N-box
U2AF2RRM2

Figure S2. Scaffold-independent analysis and titration curves for the final optimal binding motifs for each KH domain (related to Figure 2E-I, 3C, F)

(A) Schematic workflow of the NMR-based scaffold-independent analysis (SIA) [S1]. SIA reports on the nucleic acid binding specificity of a given RNA-binding protein (RBP) at each position of a nucleic acid target. Sixteen 5-mer DNA pools with one specific nucleotide fixed at one position, otherwise randomized, are individually titrated to each KH domain. The observed changes in chemical shift of the selected peaks are averaged and normalized for each DNA pool to obtain a score for the nucleotide position and type preference.

(B) Comparisons of chemical shift perturbations (CSPs) of all four FUBP1 KH domains upon addition of the optimal nucleotide motifs as either DNA or RNA (1:1 molar ratio of protein to RNA). This shows that DNA and RNA binding are very similar for all four KH domains of FUBP1.

(C) NMR titration and dissociation constant (K_D) calculation for the binding of FUBP1 KH1 (100 μM) with TTTTG up to a protein/DNA molar ratio of 1:8. The indicated residues are used for the calculation of K_D. As expected for interactions of KH domains to nucleic acids, the changes in chemical shift are mostly mapped to the α1 and α2 helices and the GXXG loop.

(D) NMR titration and K_D calculation for the binding of FUBP1 KH2 (100 μM) with TTTGT up to a protein/DNA molar ratio of 1:16. The marked residues are used for the calculation of K_D. As expected, the changes in chemical shift are mostly mapped to the α1 and α2 helices and the GXXG loop.

(E) NMR titration and K_D calculation for the binding of FUBP1 KH3 (100 μM) with TCTGT up to a protein/DNA molar ratio of 1:12. The marked residues are used for the calculation of K_D. As expected, the changes in chemical shift are mostly mapped to the α1 and α2 helices and the GXXG loop.

(F) NMR titration and K_D calculation for the binding of FUBP1 KH4 (100 μM) with TTTTG up to a protein/DNA molar ratio of 1:12. The marked residues are used for the calculation of K_D. As expected, the changes in chemical shift are mostly mapped to the α1 and α2 helices and the GXXG loop.

(G) NMR titration and ITC of FUBP1 KH1–2 binding to a DNA oligonucleotide containing an optimal DNA motif for each KH domain derived by SIA linked by an A4 linker (TTTGTAAAATTTTG). Consistent with the K_D values from ITC, the NMR titration indicates binding in an intermediate exchange regime.

(H) NMR titration and ITC of FUBP1 KH2-3 binding to a DNA oligonucleotide containing an optimal DNA motif for each KH domain derived by SIA linked by an A4 linker (TCTGTAAAATTTGT). Consistent with the K_D values from ITC, the NMR titration indicates binding in an intermediate exchange regime.

(I) NMR titration and ITC of FUBP1 KH3–4 binding to a DNA oligonucleotide containing an optimal DNA motif for each KH domain derived by SIA linked by an A4 linker (TTTTGAAAATCTGT). Consistent with the K_D values from ITC, the NMR titrations indicate binding in an intermediate exchange regime.

(J) Motif enrichment in the *in vivo* FUBP1 iCLIP data. Disjunct 4-mer frequencies were calculated in extended windows (5-nt binding site ± 5 nt) around the top 20% of binding sites based on expression-normalized iCLIP signal and in non-bound regions in the same introns excluding a 20-nt region downstream of the 5' ss and a 150-nt region upstream of the branch point (BP). Enrichment for each motif is defined as the distance for each data point to the diagonal in the scatterplot of relative motif frequencies of the top 20% vs bottom 20% of binding sites.

(K) Positional enrichment of FUBP1 binding motifs and control motifs relative to the BP. UUU+A/G/C, that is, 4-mers containing UUU interspersed at any position with A/G/C. Control motif sets are mononucleotide tracts interspersed by one other nucleotide. 4-mer frequencies were calculated position-wise upstream of the BP and compared to the average 4-mer frequencies in an

intronic control region (a 100-nt-long region 100 nt downstream of the 5' splice site). Shaded regions correspond to the main binding regions of FUBP1 (red) and SF3B1 (blue).

(L) Abundance of FUBP1 binding motifs (UUU+A/G) at 3' ss of human introns. Abundance for other mononucleotide motifs (AAA + C/G & AAAA, CCC + G/C & CCCC, GGG + A/C & GGGG) is given for the purpose of comparison.

(M) Total luminescence and fluorescence measurements were used to estimate the amounts of FUBP1 or the mutants $FUBP1^{A38D}$, $FUBP1^{\Delta C}$, and $FUBP1^{W586,615R}$ paired with wild-type U2AF2 and SF1 (orange), BCL2L1-BAD as a positive control pair (green) and pairs that are not known to interact with each other as negative controls (gray) in bioluminescence resonance energy transfer (BRET)-based assay. Acceptor/donor ratios are similar for all pairs, making the cBRET values more comparable to each other.

(N) NMR titration of $U2AF2^{RRM2}$ with $FUBP1^{N\text{-}box}$ up to sixfold molar excess (left). Significantly shifted peaks are enlarged. The peaks with a chemical shift perturbation (CSP) ≥ 0.1 are shown in red along with corresponding residues on the structure of $U2AF2^{RRM2}$ (right) (PDB ID: 8P25).

Figure S3

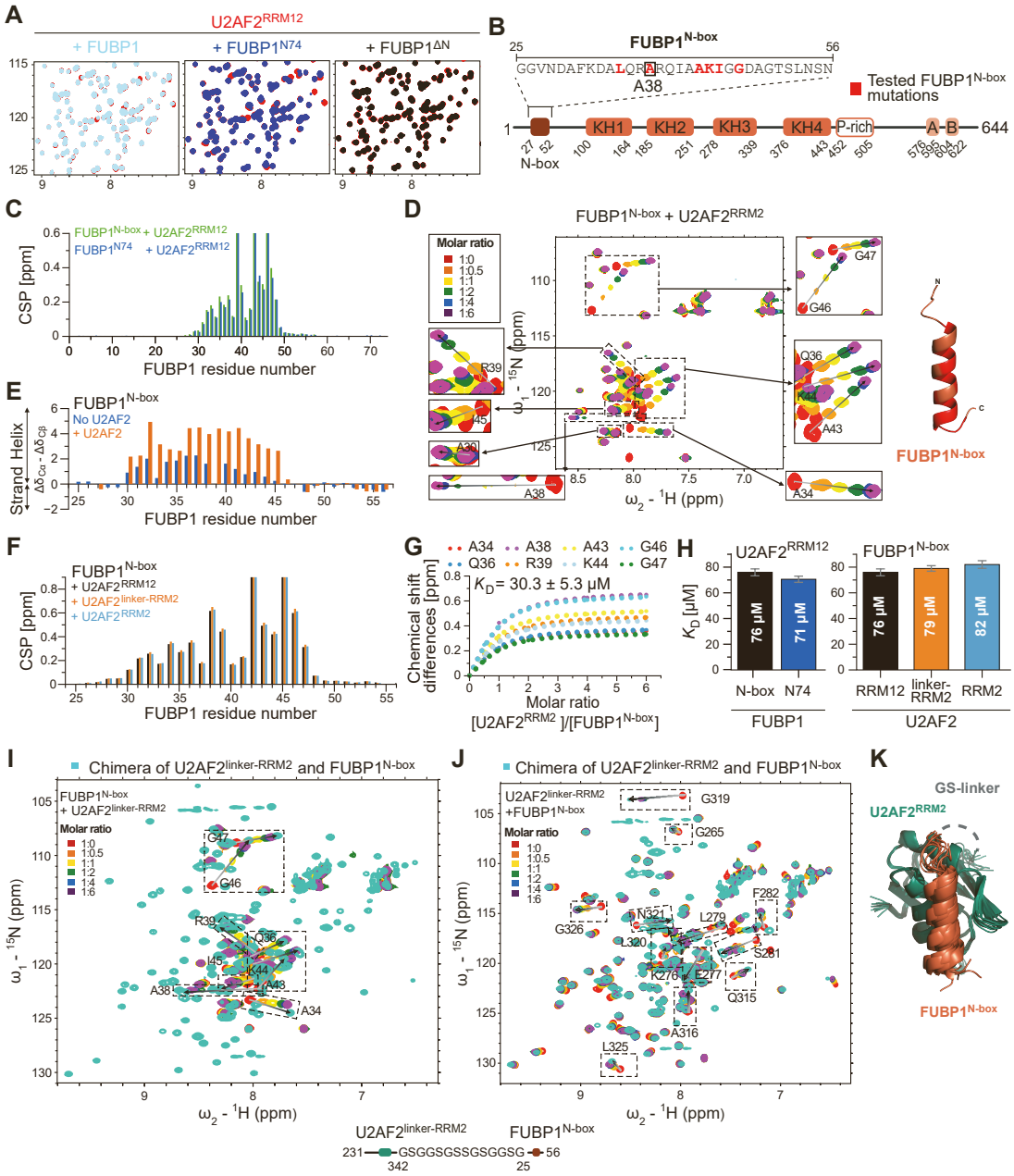


Figure S3. Determination of the minimal interaction interface between FUBP1$^{N\text{-}box}$ and U2AF2^{RRM2} (related to Figure 3F-H)

(A) Comparison of a selected region of the one-point NMR titrations (0.5 molar ratio) of U2AF2^{RRM2} (red) with full-length FUBP1 (cyan), FUBP1^{N74} (blue), and FUBP1$^{\Delta N}$ (black) showing significant chemical shift changes.

(B) Overview of the FUBP1 construct used for NMR and BRET experiments. Red color marks mutations that are tested for effect on binding of FUBP1 with U2AF2.

(C) Comparison of CSPs upon the titration of a shortened FUBP1 N-terminal construct ($FUBP1^{N\text{-}box}$, aa 25–56; green) and $FUBP1^{N74}$ (blue) with $U2AF2^{RRM12}$.

(D) NMR titration of $FUBP1^{N\text{-}box}$ with $U2AF2^{RRM2}$ up to sixfold molar excess (left). Significantly shifted peaks are boxed and enlarged. The peaks with a CSP ≥ 0.1 are highlighted in red along with corresponding residues on the structure of $FUBP1^{N\text{-}box}$ (right) (PDB ID: 8P25).

(E) Comparison of the C_α and C_β chemical shift-derived secondary structure of free $FUBP1^{N\text{-}box}$ (blue) and $FUBP1^{N\text{-}box}$ bound to $U2AF2^{RRM2}$ (orange). The fractional helical conformation for residues 30–45 in the absence of U2AF2 is further increased upon binding to U2AF2.

(F) Comparison of the CSP of $FUBP1^{N\text{-}box}$ titrations with U2AF2 constructs of various lengths ($U2AF2^{RRM12}$, black; $U2AF2^{linker\text{-}RRM2}$, orange; $U2AF2^{RRM2}$, light blue).

(G) Calculation of K_D for the $FUBP1^{N\text{-}box}$ and $U2AF2^{RRM2}$ interaction derived by NMR titration. The changes in chemical shift of selected residues in the titration of $FUBP1^{N\text{-}box}$ with $U2AF2^{RRM2}$ (shown in panel D) are plotted against the molar ratio of ligand to titrant.

(H) Comparison of KD values for the interaction of $FUBP1^{N\text{-}box}$ (black) and $FUBP1^{N74}$ (blue) with $U2AF2^{RRM12}$ and $FUBP1^{N\text{-}box}$ with $U2AF2^{RRM12}$ (black), $U2AF2^{linker\text{-}RRM2}$ (orange), and $U2AF2^{RRM2}$ (light blue), determined by ITC (**Table S2**). The measurements were performed in triplicates and data are represented as mean ± SD.

(I) Overlay of the ^{1}H–^{15}N HSQC spectra of the chimeric construct $U2AF2^{linker\text{-}RRM2}$/$FUBP1^{N\text{-}box}$ (cyan) and $FUBP1^{N\text{-}box}$ titrated with $U2AF2^{linker\text{-}RRM2}$ (molar ratio of 1:6).

(J) Overlay of the ^{1}H–^{15}N HSQC spectra of the chimeric construct $U2AF2^{linker\text{-}RRM2}$/$FUBP1^{N\text{-}box}$ (cyan) and $U2AF2^{linker\text{-}RRM2}$ titrated with $FUBP1^{N\text{-}box}$ (molar ratio of 1:6).

(K) NMR ensemble (10 lowest energy structures) of the chimeric construct $U2AF2^{linker\text{-}RRM2}$ (green)/$FUBP1^{N\text{-}box}$ (brown). The end of the flexible linker between RRM1-RRM2 (231–245) is not shown, the artificial GS-linker between the C terminus of U2AF2 RRM2 and the N-terminal region of $FUBP1^{N\text{-}box}$ are indicated by gray dashed lines (PDB ID: 8P25).

Figure S4

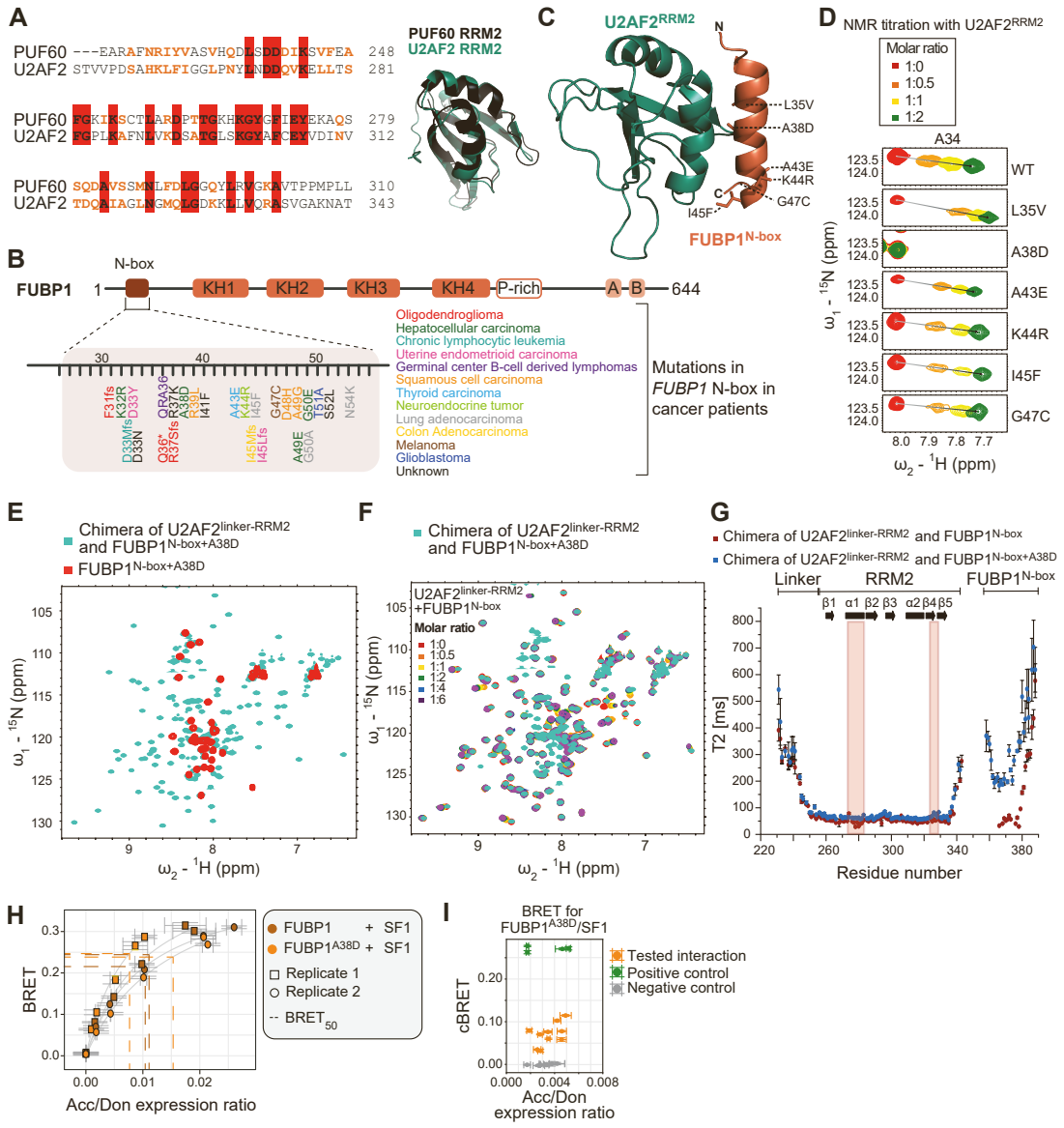


Figure S4. The effects of cancer-related mutations in the FUBP1 N-box on the interaction with U2AF2 RRM2 (related to Figure 3C, 3I-J)

(A) Sequential alignment (Clustal Omega [S4]) of the RRM2 domains of human PUF60 and U2AF2, mapping conserved residues (red) and similar residues (orange). Overlay of the structure of PUF60 RRM2 (black) and U2AF2 RRM2 (green) (adapted from [S2] and [S3]; PDB IDs: 2KXH, 6TR0).

(B) FUBP1 N-box mutations identified in different cancer types. Databases (see STAR Methods) were screened for the occurrence of cancer-related mutations within the region of *FUBP1* encoding for the N-box, yielding one insertion, five frameshifts (fs) leading to a premature termination codon (*) and 20 missense variants.

(C) Cancer-related mutations (labeled and side chains shown on the calculated structure of a chimeric construct of U2AF2^{RRM2} and FUBP1$^{N\text{-}box}$, PDB ID: 8P25) within the helical binding region of FUBP1$^{N\text{-}box}$ and located at the interfaces with U2AF2^{RRM2} were selected for further NMR study.

(D) Comparison of the changes in chemical shift of residue A34 for the titration of FUBP1$^{N\text{-}box}$ wild-type and mutants (L35V, A38D, A43E, K44R, I45F, G47C) upon adding U2AF2^{RRM2}.

(E) Overlay of the ^{1}H–^{15}N HSQC spectra of the A38D mutant of FUBP1$^{N\text{-}box}$ (red) with the chimeric construct U2AF2$^{linker\text{-}RRM2}$/FUBP1$^{N\text{-}box}$ with A38D mutation (cyan).

(F) Overlay of the ^{1}H–^{15}N HSQC spectra of the chimeric construct U2AF2$^{linker\text{-}RRM2}$/FUBP1$^{N\text{-}box}$ with A38D mutation (cyan) with the titration of FUBP1$^{N\text{-}box}$ with U2AF2$^{linker\text{-}RRM2}$ shows that the mutant spectrum resembles those of the unbound individual components.

(G) Comparison of the ^{15}N T_2 relaxation rates of the chimeric constructs U2AF2$^{linker\text{-}RRM2}$/FUBP1$^{N\text{-}box}$ wild-type and A38D mutant. Increased T_2 relaxation rates in the N-box helix of the A38D mutant chimera compared to the wild-type is consistent with much weaker binding of the mutant to the U2AF2 RRM2.

(H) BRET titration curves shown for FUBP1 and FUBP1^{A38D} versus SF1. As expected, mutation of the FUBP1 N-box does not result in significant loss of binding to SF1. Two biological replicates are shown, each done in technical triplicates. Error bars represent the standard deviation.

(I) Total luminescence (Don) and fluorescence (Acc) ratios were determined for FUBP1 and FUBP1^{A38D} versus SF1. Acceptor/donor ratios are similar for all pairs making the cBRET values more comparable to each other.

Figure S5

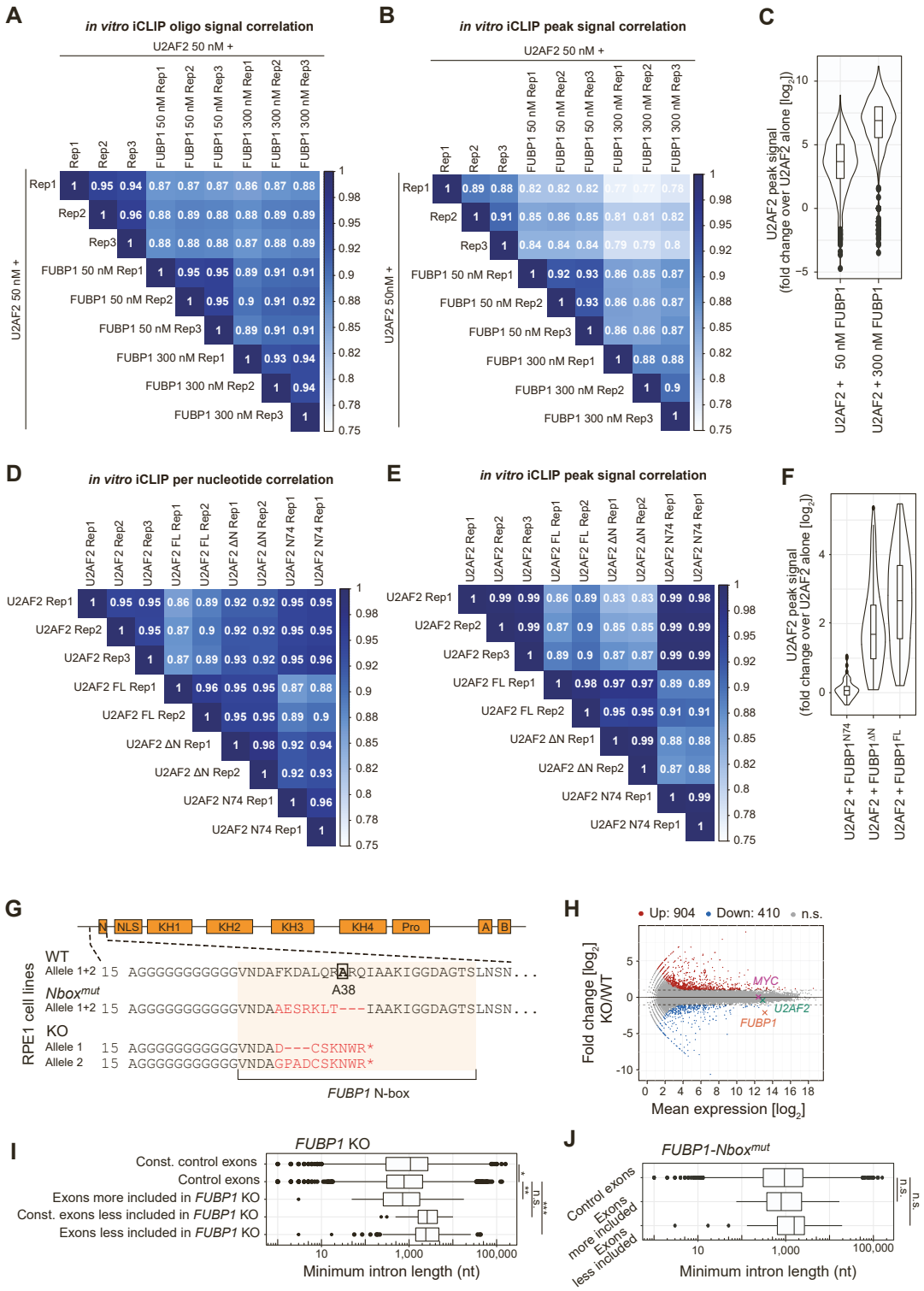

A
in vitro iCLIP oligo signal correlation
U2AF2 50 nM +
U2AF2 50 nM +
Rep1
Rep2
Rep3
FUBP1 50 nM Rep1
FUBP1 50 nM Rep2
FUBP1 50 nM Rep3
FUBP1 300 nM Rep1
FUBP1 300 nM Rep2
FUBP1 300 nM Rep3
B
in vitro iCLIP peak signal correlation
U2AF2 50 nM +
U2AF2 50nM +
C
U2AF2 peak signal
(fold change over U2AF2 alone [log2])
U2AF2 + 50 nM FUBP1
U2AF2 + 300 nM FUBP1
D
in vitro iCLIP per nucleotide correlation
U2AF2 Rep1
U2AF2 Rep2
U2AF2 Rep3
U2AF2 FL Rep1
U2AF2 FL Rep2
U2AF2 ΔN Rep1
U2AF2 ΔN Rep2
U2AF2 N74 Rep1
E
in vitro iCLIP peak signal correlation
F
U2AF2 peak signal
(fold change over U2AF2 alone [log2])
U2AF2 + FUBP1N74
U2AF2 + FUBP1ΔN
U2AF2 + FUBP1FL
G
N
NLS
KH1
KH2
KH3
KH4
Pro
A
B
RPE1 cell lines
WT
Allele 1+2 15 AGGGGGGGGGGGVNDAFKDALQHAQIAAKIGGDAGTSLNSN...
Nbox^mut
A38
Allele 1+2 15 AGGGGGGGGGGGVNDAAESRKLT---IAAKIGGDAGTSLNSN...
KO
Allele 1 15 AGGGGGGGGGGGVNDAD---CSKNWR*
Allele 2 15 AGGGGGGGGGGGVNDAGPADCSKNWR*
FUBP1 N-box
H
Up: 904
Down: 410
n.s.
Fold change [log2] KO/WT
MYC
U2AF2
FUBP1
Mean expression [log2]
I
FUBP1 KO
Const. control exons
Control exons
Exons more included in FUBP1 KO
Const. exons less included in FUBP1 KO
Exons less included in FUBP1 KO
n.s.
Minimum intron length (nt)
J
FUBP1-Nbox^mut
Control exons
Exons more included
Exons less included
n.s.
Minimum intron length (nt)

Figure S5. Reproducibility between replicates and changes in U2AF2^{RRM12} binding from *in vitro* iCLIP experiments and expression and splicing changes upon *FUBP1* KO (related to Figure 4A-F, 5A-B)

(A) Reproducibility of *in vitro* iCLIP data with oligonucleotide-derived transcript library. The correlation matrix shows pairwise Pearson correlation of U2AF2^{RRM12} crosslink events per oligonucleotide (n = 1,998) between samples. Experiments were performed with U2AF2^{RRM12} alone (50 nM) and with the addition of full-length FUBP1 at 50 or 300 nM.

(B) Reproducibility of *in vitro* iCLIP data with oligonucleotide-derived transcript library. The correlation matrix shows pairwise Pearson correlation of total U2AF2^{RRM12} crosslink events inside U2AF2 binding sites between samples (1,831 oligonucleotides harbor a U2AF2 binding sites according to U2AF2 *in vivo* iCLIP). Experiments as in panel A.

(C) Comparative boxplot of normalized U2AF2^{RRM12} crosslink events per binding site between conditions (n = 1,504). Experiments as in panel A.

(D) Reproducibility of *in vitro* iCLIP data with eight long *in vitro* transcripts [S5]. The correlation matrix shows pairwise Pearson correlation of U2AF2^{RRM12} crosslink events per nucleotide over all *in vitro* transcripts between samples. Experiments were performed with U2AF2^{RRM12} alone (50 nM) and with the addition of full-length FUBP1^{FL}, FUBP1^{N74}, and FUBP1$^{\Delta N}$ (all 50 nM).

(E) Reproducibility of *in vitro* iCLIP data with eight long *in vitro* transcripts [S5]. Correlation matrix shows pairwise Pearson correlation of total binding signals (n = 109) between samples. Experiments as in panel D.

(F) Comparative boxplot of normalized U2AF2^{RRM12} crosslink events between conditions (n = 109). Experiments as in panel D.

(G) Zoom-in of the FUBP1$^{N\text{-}box}$ sequence, which when targeted with CRISPR/Cas9 results in a knockout cell line (*FUBP1* KO) and a mutant cell line (*FUBP1-Nbox^{mut}*), in which FUBP1 lacks the U2AF2 interaction surface.

(H) Log_2 fold change versus mean expression for genes upon *FUBP1* KO in RPE1 cells.

(I) Minimum adjacent intron length for cassette exons that are more or less included and for constitutive exons less included in *FUBP1-Nbox^{mut}* RPE1 cells (n = 123/249/27) compared to unchanged control exons (n = 4,584) and unchanged constitutive control exons (n = 5,717).

(J) Minimum adjacent intron length for cassette exons that are more or less included in *FUBP1-Nbox^{mut}* RPE1 cells (n = 36/45) compared to unchanged control exons (n = 10,678).

Figure S6

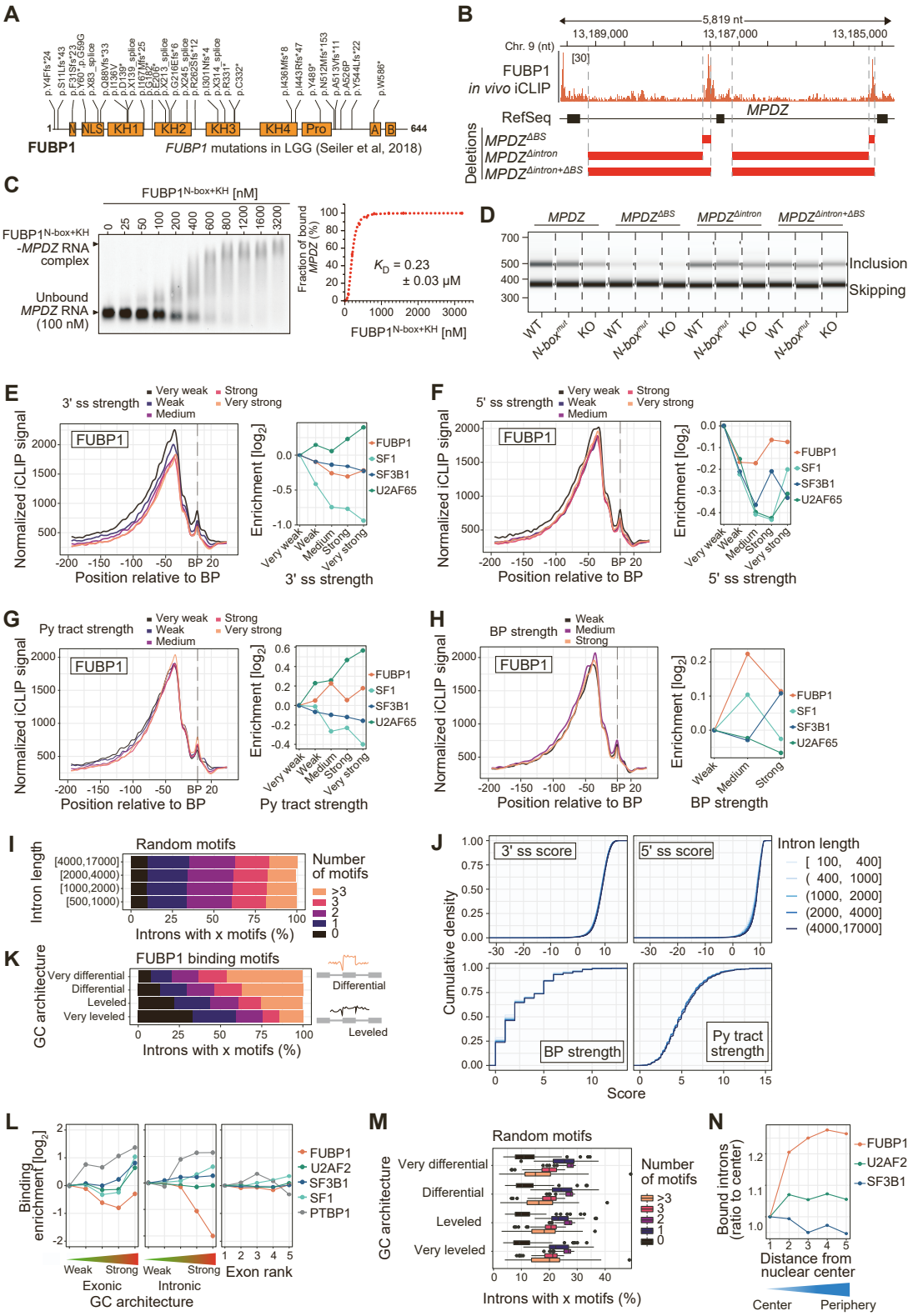

A
FUBP1
FUBP1 mutations in LGG (Seiler et al, 2018)
B
5,819 nt
13,189,000
13,187,000
13,185,000
Chr. 9 (nt)
FUBP1
in vivo iCLIP
RefSeq
MPDZ
Deletions
C
FUBP1N-box+KH [nM]
FUBP1N-box+KH -MPDZ RNA complex
Unbound MPDZ RNA (100 nM)
Fraction of bound MPDZ (%)
KD = 0.23 ± 0.03 μM
D
MPDZ
MPDZΔBS
MPDZΔintron
MPDZΔintron+ΔBS
Inclusion
Skipping
WT
N-boxmut
KO
E
3' ss strength
Very weak
Weak
Medium
Strong
Very strong
Normalized iCLIP signal
FUBP1
Position relative to BP
Enrichment [log2]
SF1
SF3B1
U2AF65
F
5' ss strength
G
Py tract strength
H
BP strength
I
Random motifs
Intron length
Number of motifs
Introns with x motifs (%)
J
3' ss score
5' ss score
Intron length
Cumulative density
BP strength
Py tract strength
Score
K
FUBP1 binding motifs
GC architecture
Very differential
Differential
Leveled
Very leveled
L
Binding enrichment [log2]
U2AF2
PTBP1
Exonic
Intronic
Exon rank
M
N
Bound introns (ratio to center)
Distance from nuclear center
Center
Periphery

Figure S6. FUBP1 effects on long introns (related to Figure 5C-H)

(A) Position and identity of FUBP1 loss-of-function (LoF) mutations in glioma patients with 1p/19q deletion-positive background [S6].

(B) Genome browser view of the region included in the *MPDZ* minigene displaying the *in vivo* iCLIP data (crosslink events per nucleotide) of FUBP1 (orange). Deletions of introns with/without FUBP1 binding sites are indicated below with red bars.

(C) EMSA experiment to demonstrate binding of recombinant $FUBP1^{N\text{-}box+KH}$ (aa 1–457, 25–3200 nM) to a fluorescently labeled 132-nt RNA fragment from *MPDZ* (100 nM). Agarose gel image (bottom) and quantification (top) with fitted curve show FUBP1–RNA binding in a nanomolar range (K_D = 0.23 ± 0.03 μM).

(D) Capillary electrophoresis of exon inclusion levels upon intron shortening in the *MPDZ* minigene.

(E) Metaprofile showing the number of crosslink events of FUBP1 relative to the branch point in dependency on 3' splice site strength. iCLIP signals are normalized for expression and then averaged per nucleotide over all introns (left). Binding enrichment quantification: Area under the curve (AUC) in each intron class compared to the AUC in introns with very low 3' splice site strength (right).

(F) Metaprofile showing the number of crosslink events of FUBP1 relative to the branch point in dependency on 5' splice site strength. iCLIP signals are normalized for expression and then averaged per nucleotide over all introns (left). Binding enrichment quantification: AUC in each intron class compared to the AUC in introns with very low 5' splice site strength (right).

(G) Metaprofile showing the number of crosslink events of FUBP1 relative to the branch point in dependency on Py tract strength. iCLIP signals are normalized for expression and then averaged per nucleotide over all introns (left). Binding enrichment quantification: AUC in each intron class compared to the AUC in introns with very low Py tract strength (right).

(H) Metaprofile showing the number of crosslink events of FUBP1 relative to the branch point in dependency on BP strength. iCLIP signals are normalized for expression and then averaged per nucleotide over all introns (left). Binding enrichment quantification: AUC in each intron class compared to the AUC in introns with very weak BP strength (right).

(I) Fraction of introns with 0, 1, 2, 3, or > 3 motif sets of size 9 of random 4-mers in dependency on intron length. Random sets were drawn 100 times and the resulting fractions were then averaged.

(J) Cumulative distribution of splice site features conditioned on intron length.

(K) Number of FUBP1-binding motifs upstream of the BP ([−100 nt; −26 nt]) in dependency on differential GC content. Differential GC content is the GC content of the exon minus that of the first 100 nt of the downstream intron.

(L) Enrichment of FUBP1 binding upstream of the branch point in dependency on exon/intron GC content and exon rank. In the underlying metaprofiles, iCLIP signals are normalized for expression and then averaged per nucleotide over all introns.

(M) Fraction of introns with 0, 1, 2, 3 or > 3 motif sets of size 9 of random 4-mers in dependency on differential GC content. Random sets are drawn 100 times and resulting fractions are then averaged.

(N) Percent of introns bounds through different scopes of Euclidean distances where 1 means the nuclear center and 5 is the periphery. Enrichment is shown compared to the first scope. Based on data from [S7].

Figure S7

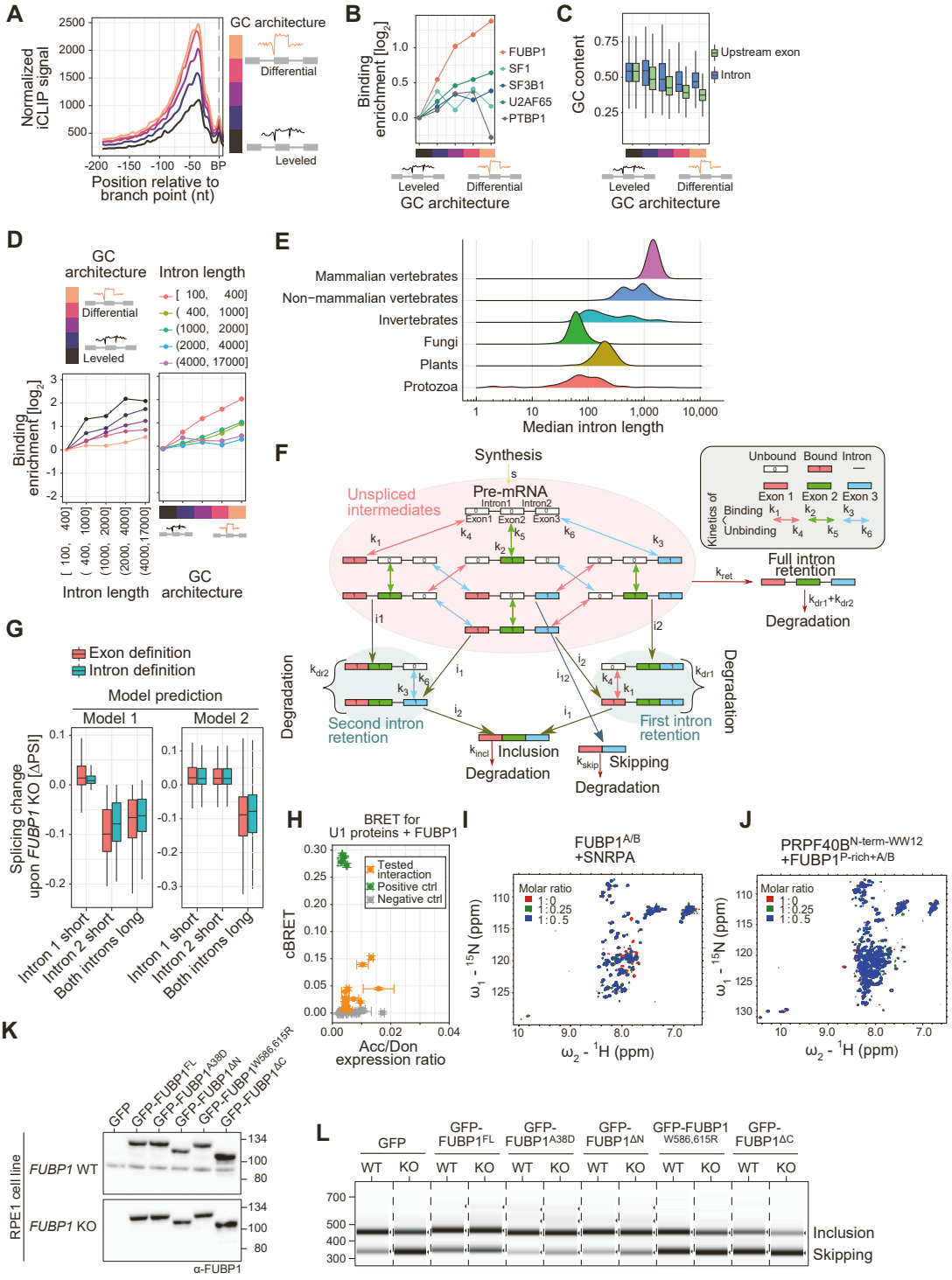

A
Normalized iCLIP signal
Position relative to branch point (nt)
GC architecture
Differential
Leveled
B
Binding enrichment [log2]
FUBP1
SF1
SF3B1
U2AF65
PTBP1
GC architecture
C
GC content
Upstream exon
Intron
D
Intron length
[100, 400]
(400, 1000]
(1000, 2000]
(2000, 4000]
(4000, 17000]
E
Mammalian vertebrates
Non−mammalian vertebrates
Invertebrates
Fungi
Plants
Protozoa
Median intron length
F
Synthesis
Pre-mRNA
Unspliced intermediates
Full intron retention
Degradation
Second intron retention
First intron retention
Inclusion
Skipping
Kinetics of
Binding
Unbinding
G
Exon definition
Intron definition
Model prediction
Model 1
Model 2
Splicing change upon FUBP1 KO [ΔPSI]
Intron 1 short
Intron 2 short
Both introns long
H
BRET for U1 proteins + FUBP1
Tested interaction
Positive ctrl
Negative ctrl
cBRET
Acc/Don expression ratio
I
FUBP1A/B +SNRPA
Molar ratio
1:0
1:0.25
1:0.5
ω1 - 15N (ppm)
ω2 - 1H (ppm)
J
PRPF40BN-term-WW12 +FUBP1P-rich+A/B
K
GFP
GFP-FUBP1FL
GFP-FUBP1A38D
GFP-FUBP1ΔN
GFP-FUBP1W586,615R
GFP-FUBP1ΔC
RPE1 cell line
FUBP1 WT
FUBP1 KO
α-FUBP1
L
WT
KO
Inclusion
Skipping

Figure S7. Characterization and modeling of FUBP1 binding behavior (related to Figure 5E, 5I-J, 6A-B, 6G-I)

(A) Metaprofile showing the number of crosslink events of FUBP1 relative to the BP in dependency on differential GC content. iCLIP signals are normalized for expression and then averaged per nucleotide over all introns.

(B) Binding enrichment quantification: AUC in each intron class compared to the AUC in introns with leveled GC content.

(C) Comparison of exon and intron GC content for exons with increasing differential GC content architecture.

(D) Enrichment of FUBP1 binding upstream of the branch point in dependency on intron length and differential GC content. Exons were classified into each intron length groups and then split by GC content architectures (left panel) and vice versa (right panel).

(E) Intron length distribution between kingdoms. Analyses were performed for 174 mammals, 274 non-mammalian vertebrates, 277 invertebrates, 410 fungal species, 94 protozoa, and 145 plants.

(F) Detailed scheme of the mathematical model describing exon definition and splicing for a cassette exon flanked by two constitutive exons. After pre-mRNA synthesis (s), the three exons (indicated by boxes) can be cooperatively and reversibly bound by the pioneering spliceosome subunits U1 and U2 (these are not explicitly displayed in the scheme). Colorless (0) and colored (1) squares represent bound ("defined") and unbound ("undefined") exons, respectively. Red, green, and blue arrows represent binding to and dissociation from exons 1, 2, and 3, respectively, where k_1–k_3 are the corresponding rate constants of binding and k_4–k_6 the rate constants of dissociation. Based on the exon definition patterns (highlighted by red ellipse), splicing decisions towards multiple splice isoforms (inclusion, skipping, first intron retention, second intron retention, full intron retention) are made, and it is assumed that an intron can be excised if the two neighboring exons are defined. For instance, skipping of exon 2 is possible from the state 1_0_1 and occurs with the rate i_{12}. Likewise, splicing of the first intron occurs from the species P1_1_0 and P1_1_1 (rate i_1), and splicing of the second intron from P0_1_1 and P1_1_1 (rate i_2). The inclusion isoform is generated in two steps, i.e., from the subsequent removal of intron 1 and intron 2 in random order. All terminal splice products are subject to degradation (k_{incl}: degradation rate constant of inclusion, k_{skip}: skipping, k_{dr1}: first intron retention, k_{dr2}: second intron retention, $k_{dr1}+k_{dr2}$: full intron retention).

(G) The intron and exon definition models show similar splicing changes upon *FUBP1* knockout (KO). We simulated the splicing changes upon *FUBP1* KO based on the assumption that FUBP1 affects the rate of spliceosome binding to the 3' splice site of long introns (left panel) or the rate off splicing catalysis across long introns (right panel) as described in detail in the STAR Methods. To account for the heterogeneity of exons in the human genome, we randomly sampled the kinetic parameters of the model 10,000 times to generate an ensemble of 10,000 in silico exons. We then simulated *FUBP1* KO for each in silico exon, assuming that FUBP1 selectively enhances the rate of splicing for long introns, and considered three scenarios reflecting different length configurations of upstream and downstream introns (see STAR Methods for details). The boxplots show the distributions of ΔPSI = PSI(KO) – PSI(control) values for exon (red) and intron definition (blue) across all exons.

(H) Total luminescence and fluorescence measurements were used to estimate the amount of FUBP1 paired with the components of U1 complex (orange), BCL2L1–BAD as a positive control pair (green) and pairs that are not known to interact with each other as negative controls

(gray). Acceptor/donor ratios are similar for all pairs making the cBRET values more comparable to each other.

(I) $^{1}H–^{15}N$ HSQC spectra of the titration of $FUBP1^{A/B}$ with SNRPA up to a molar ratio of 1:1.

(J) $^{1}H–^{15}N$ HSQC spectra of the titration of $PRPF40B^{N\text{-}term\text{-}WW12}$ with $FUBP1^{P\text{-}rich+A/B}$ up to a molar ratio of 1:0.5.

(K) Western blot to verify FUBP1 construct expression after transfection of RPE1 WT and *FUBP1* KO cells.

(L) Capillary electrophoresis of exon inclusion levels of the *MPDZ* minigene after transfection of RPE1 WT and *FUBP1* KO cells with different FUBP1 constructs.

Supplementary Tables

Table S2. Binding affinities and stoichiometries determined by ITC experiments (related to Figures 2D, 2F, S2G–I, and S3H). Experiments were performed for different FUBP1 N-terminal constructs (FUBP1$^{N\text{-}box}$, FUBP1^{N74}) with U2AF2 constructs (U2AF2^{RRM12}, U2AF2$^{linker\text{-}RRM2}$ and U2AF2^{RRM2}) and various FUBP1 KH domain constructs (FUBP1^{KH12}, FUBP1^{KH23}, FUBP1^{KH34} and FUBP1^{KH}) with DNA or RNA.

Analyte	Titrant	N sites	K_D [μM]	Repeats
U2AF2^{RRM12}	FUBP1$^{N\text{-}box}$	0.98 ± 0.17	75.93 ± 2.70	3
U2AF2^{RRM12}	FUBP1^{N74}	0.85 ± 0.07	70.57 ± 2.41	3
U2AF2$^{linker\text{-}RRM2}$	FUBP1$^{N\text{-}box}$	0.91 ± 0.16	78.97 ± 2.16	3
U2AF2^{RRM2}	FUBP1$^{N\text{-}box}$	0.87 ± 0.16	82.03 ± 2.98	3
FUBP1^{KH12}	TTTGTAAAATTTTG	0.78 ± 0.07	4.71 ± 1.45	3
FUBP1^{KH23}	TCTGTAAAATTTGT	0.76 ± 0.09	1.15 ± 0.48	3
FUBP1^{KH34}	TTTTGAAAATCTGT	0.74 ± 0.04	0.87 ± 0.10	3
VPS13D RNA	FUBP1^{KH}	1.38 ± 0.04	0.428 ± 0.062	3

Supplementary References

[S1] Beuth, Barbara, María Flor García-Mayoral, Ian A. Taylor, and Andres Ramos. 2007. "Scaffold-Independent Analysis of RNA-Protein Interactions: The Nova-1 KH3-RNA Complex." *Journal of the American Chemical Society* 129 (33): 10205–10.

[S2] Kang, Hyun-Seo, Carolina Sánchez-Rico, Stefanie Ebersberger, F. X. Reymond Sutandy, Anke Busch, Thomas Welte, Ralf Stehle, et al. 2020. "An Autoinhibitory Intramolecular Interaction Proof-Reads RNA Recognition by the Essential Splicing Factor U2AF2." *Proceedings of the National Academy of Sciences of the United States of America* 117 (13): 7140–49.

[S3] Cukier, Cyprian D., David Hollingworth, Stephen R. Martin, Geoff Kelly, Irene Díaz-Moreno, and Andres Ramos. 2010. "Molecular Basis of FIR-Mediated c-Myc Transcriptional Control." *Nature Structural & Molecular Biology* 17 (9): 1058–64.

[S4] Madeira, Fábio, Young Mi Park, Joon Lee, Nicola Buso, Tamer Gur, Nandana Madhusoodanan, Prasad Basutkar, et al. 2019. "The EMBL-EBI Search and Sequence Analysis Tools APIs in 2019." *Nucleic Acids Research* 47 (W1): W636–41.

[S5] Sutandy, F. X. Reymond, Stefanie Ebersberger, Lu Huang, Anke Busch, Maximilian Bach, Hyun-Seo Kang, Jörg Fallmann, et al. 2018. "In Vitro iCLIP-Based Modeling Uncovers How the Splicing Factor U2AF2 Relies on Regulation by Cofactors." *Genome Research* 28 (5): 699–713.

[S6] Seiler, Michael, Shouyong Peng, Anant A. Agrawal, James Palacino, Teng Teng, Ping Zhu, Peter G. Smith, Cancer Genome Atlas Research Network, Silvia Buonamici, and Lihua Yu. 2018. "Somatic Mutational Landscape of Splicing Factor Genes and Their Functional Consequences across 33 Cancer Types." *Cell Reports* 23 (1): 282–96.e4.

[S7] Tammer, Luna, Ofir Hameiri, Ifat Keydar, Vanessa Rachel Roy, Asaf Ashkenazy-Titelman, Noélia Custódio, Itay Sason, et al. 2022. "Gene Architecture Directs Splicing Outcome in Separate Nuclear Spatial Regions." *Molecular Cell*. Elsevier.

However, I performed the cloning of the ORFs, site-directed mutagenesis and generated constructs in the low-throughput (eppi tube format). Next, I adapted these steps in medium-throughput (plate format). The protocol is described in Appendix, 5.1. The final pipeline followed by BRET-assay was tested in the second collaborative project **see Article II**.

2.4 Article II: Systematic discovery of protein interaction interfaces using AlphaFold and experimental validation

Summary

This project is focused on benchmarking AlphaFold-Multimer (AF-MM), its metrics and the application to identify novel protein interaction interfaces followed by experimental validation.

AF-MM is a machine learning-based tool to predict structures of protein interactions and complexes. While this tool was tested to predict different PPI interface types, there is a general lack of a comprehensive assessment of sensitivity and specificity and the potential biases of the tool and its metrics. The benchmarking of AF-MM is essential for its application in the prediction of PPI interfaces. Therefore, we systematically benchmarked the tool's ability to predict domain-domain and domain-motif interfaces. The predicted models were compared to the solved structures and found that 35 % of the putative DMIs were predicted correctly including the positions of sidechains, whereas 34 % had correct backbone predictions.

We also evaluated the metrics of AF-MM for their application in distinguishing known DMIs from random DMI pairs. We also examined the effect of sequence length on the tool's performance and found that long fragments of full-length proteins might worsen the predictions.

These findings motivated us to develop a fragmentation approach, where the overlapping fragments were used to predict novel DMIs in human PPIs. We applied this strategy to 62 PPIs from the HuRI dataset, where proteins are disease-associated. This strategy improved the sensitivity but decreased AF-MM's specificity. We further manually inspected high-scoring models. We selected some models for further experimental validation. Using a plate-based bioluminescence resonance energy transfer (BRET) assay, known for its sensitivity in detecting point mutation effects and motif-mediated protein-protein interactions (PPIs), we tested 28 of the 62 PPIs, where BRET signals were significant

for 11 of these 28 PPIs. Using the putative structures we selected key interacting residues, that are also conserved and designed mutations that potentially can disrupt the predicted interface and deletions of the predicted motif. We further validated seven predicted interfaces. Moreover, we discovered a novel interface between PEX3 and PEX16 and proposed a model for their interaction with PEX19. However, our experimental data also showed inaccuracies and limitations of AF predictions, particularly for FBXO28-STX1B, STX1B-VAMP2, ESRRG-PSMC5 and TRIM37-PNKP interfaces, which need more studies for interface elucidation.

In summary, this project provided a thorough assessment of AF-MM and its metrics, a protein fragmentation strategy predicting novel PPI interfaces, successfully applied to proteins likely associated with neurodevelopmental disorders. Our prediction, experimentally validated for 6/7 novel interfaces offers molecular insights, while also highlighting the potential limitations of AF-MM and the need for further advancements to increase prediction accuracy. So far, this is the largest effort in using AF-MM for PPI interface prediction coupled with experimentally validating predicted interfaces.

Statement of contribution

This is a collaborative project in which I participated in the following aspects: data curation, cloning open reading frames (ORFs), designing mutations, site-directed mutagenesis, and experimentally validating the putative interfaces predicted by AF-MM. I also participated in organizing and analyzing the experimental data. Additionally, I participated in writing the methods section and revising the manuscript.

Supervisor confirmation

Article

SOURCE DATA | TRANSPARENT PROCESS | OPEN ACCESS | CHECK FOR UPDATES

molecular systems biology

Systematic discovery of protein interaction interfaces using AlphaFold and experimental validation

Chop Yan Lee[1,5], Dalmira Hubrich[1,5], Julia K Varga[2,5], Christian Schäfer[1], Mareen Welzel[1], Eric Schumbera[1,4], Milena Djokic[1], Joelle M Strom[1], Jonas Schönfeld[1], Johanna L Geist[1], Feyza Polat[1], Toby J Gibson[3], Claudia Isabelle Keller Valsecchi[1], Manjeet Kumar[3], Ora Schueler-Furman[2]✉ & Katja Luck[1]✉

Abstract

Structural resolution of protein interactions enables mechanistic and functional studies as well as interpretation of disease variants. However, structural data is still missing for most protein interactions because we lack computational and experimental tools at scale. This is particularly true for interactions mediated by short linear motifs occurring in disordered regions of proteins. We find that AlphaFold-Multimer predicts with high sensitivity but limited specificity structures of domain-motif interactions when using small protein fragments as input. Sensitivity decreased substantially when using long protein fragments or full length proteins. We delineated a protein fragmentation strategy particularly suited for the prediction of domain-motif interfaces and applied it to interactions between human proteins associated with neurodevelopmental disorders. This enabled the prediction of highly confident and likely disease-related novel interfaces, which we further experimentally corroborated for FBXO23-STX1B, STX1B-VAMP2, ESRRG-PSMC5, PEX3-PEX19, PEX3-PEX16, and SNRPB-GIGYF1 providing novel molecular insights for diverse biological processes. Our work highlights exciting perspectives, but also reveals clear limitations and the need for future developments to maximize the power of Alphafold-Multimer for interface predictions.

Keywords AlphaFold; Protein Interaction Interface Prediction; Linear Motifs; Benchmarking; Experimental Validation
Subject Categories Computational Biology; Structural Biology
https://doi.org/10.1038/s44320-023-00005-6
Received 3 August 2023; Revised 4 December 2023; Accepted 5 December 2023
Published online: 15 January 2024

Introduction

Protein-protein interactions (PPIs) are essential for the proper functioning of essentially all cellular processes. The last decade has seen tremendous progress in the systematic mapping of human protein interactions enabling gene function prediction and the study of genotype-to-phenotype relationships (Luck et al, 2020; Drew et al, 2017; Huttlin et al, 2021). However, to understand the molecular function of individual PPIs, co-existence or mutual exclusivity of partner proteins in protein complexes, and the effect of mutations on protein function, structural information on how these proteins interact with each other is required. Unfortunately, a structure at atomic resolution is only available for ~4% of known human PPIs (Luck et al, 2020). Modular proteins interact with each other using a variety of different functional elements such as stably folded domains, intrinsically disordered polypeptide regions, short linear motifs (hereafter referred to as motifs), or coiled-coil helices forming domain-domain, domain-motif, disorder-disorder, or coiled-coil interfaces for example. Resources such as 3did (Mosca et al, 2014) or the ELM database (ELM DB) (Kumar et al, 2022) collect observed contacts between domain types and between domains and motifs, respectively. Such interface type collections can be used to predict occurrences of known interface types in protein interactions (Weatheritt et al, 2012; Mosca et al, 2013). However, it is reasonable to expect that many more protein interface types remain to be discovered. This is likely particularly true for motif-mediated PPIs, which are anticipated to number in the hundreds of thousands or millions (Tompa et al, 2014). Motifs are short stretches of amino acids in disordered regions of proteins that usually adopt a more rigid structure upon binding to folded domains in interaction partners (Davey et al, 2012). Motif-mediated interactions are of moderate binding affinity and thus, are particularly suited to mediate dynamic cell regulatory and signaling events (Van Roey et al, 2012). However, due to the transient nature of their interactions and the disorderliness of motif-containing proteins, this mode of binding is also expected to be highly understudied. Systematically generated human protein interactome maps (Luck et al, 2020; Huttlin et al, 2021) are likely a treasure trove for the discovery of novel interface types, yet no good experimental or computational methods exist to systematically map or predict protein interaction interfaces at scale.

[1]Institute of Molecular Biology (IMB) gGmbH, 55128 Mainz, Germany. [2]Department of Microbiology and Molecular Genetics, Institute for Biomedical Research Israel-Canada, Faculty of Medicine, The Hebrew University of Jerusalem, Jerusalem 9112001, Israel. [3]Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg 69117, Germany. [4]Present address: Computational Biology and Data Mining Group Biozentrum I, 55128 Mainz, Germany. [5]These authors contributed equally: Chop Yan Lee, Dalmira Hubrich, Julia K Varga. ✉E-mail: ora.furman-schueler@mail.huji.ac.il; k.luck@imb-mainz.de

The release of the neural network-based software AlphaFold (AF) was not only a breakthrough for the prediction of monomeric structures of proteins (Jumper et al, 2021) but multiple studies published shortly thereafter also suggested the ability of AF to predict structures of pairwise protein interactions and complexes. Sensitivities of around 70% were reported using benchmark datasets of structurally resolved protein interactions originally developed to evaluate docking methods (Akdel et al, 2022; Bryant et al, 2022; Johansson-Åkhe et al, 2021; preprint:Evans et al, 2021). Other studies focused on structures of domain-motif interfaces to specifically evaluate AF's ability to predict structures for this mode of binding, reporting similar success rates (Akdel et al, 2022; Johansson-Åkhe et al, 2021; Tsaban et al, 2022). Only a few studies have also evaluated AF's specificity for the prediction of interface structures using controls such as random protein pairs or mutation of motifs to poly-alanine stretches (Akdel et al, 2022; Johansson-Åkhe et al, 2021; Tsaban et al, 2022). Different benchmarking studies used different versions of AF and reported on different metrics for their ability to distinguish good from bad structural models (Bryant et al, 2022; O'Reilly et al, 2023; Tsaban et al, 2022; preprint:Evans et al, 2021; Teufel et al, 2023). We generally lack a comprehensive assessment of the latest AF releases and metrics across different types of PPI interfaces for their sensitivity, specificity, and potential biases for the prediction of complex structures.

In a landmark study, researchers applied AF onto 65,000 human PPIs derived from a yeast two-hybrid-based interactome map (hereafter referred to as HuRI) and highly confident co-complex associations to structurally annotate the human interactome with AF-derived models. High confidence models were obtained for about 3000 PPIs (Burke et al, 2023). The authors noted a smaller fraction of highly confident structural models obtained for PPIs from the HuRI dataset compared to the co-complex dataset and reported that proteins in HuRI contain more intrinsic disorder and are less conserved compared to proteins from co-complex datasets. AF model confidence scores also increased for PPIs with proteins that are less disordered and more conserved, indicating that AF predictions work less well for PPIs mediated by interfaces involving disordered regions such as domain-motif interfaces, which likely dominate the human interactome (Tompa et al, 2014). However, AF benchmarking studies reported similarly high success rates for domain-motif interfaces compared to general docking benchmark datasets (Tsaban et al, 2022; Akdel et al, 2022). These discrepancies in sensitivities could be a result of two possible factors. First, they might point to differences in AF performance if small interacting fragments are used for interface prediction, as done in the benchmark studies, versus full length sequences used for structure prediction in (Burke et al, 2023). Second, these discrepancies could also point to difficulties of AF to predict structures of interface types involving disordered regions that have not been solved before, of which there are likely many in HuRI. It remains to be addressed to what extent these two possible factors contribute to the challenges encountered specifically for domain-motif interface modeling.

Determination of accuracies of novel predicted interface structures by AF ultimately requires experimentation. AF interface predictions for individual PPIs have occasionally been experimentally corroborated (Mishra et al, 2023; Bronkhorst et al, 2023). A more systematic experimental confirmation of AF interface models has been conducted using crosslinking mass spectrometry (XL-MS) (Burke et al, 2023; O'Reilly et al, 2023). While in-cell XL-MS is a very elegant approach to obtain experimental information on PPI interfaces in unperturbed settings, it is still a method that is only accessible to few experts in the field. Other experimental approaches are needed, which can, ideally at high throughput, confirm predicted interfaces for PPIs. In this study, we thoroughly benchmarked the two most recent versions of AlphaFold-Multimer (hereafter referred to as AF) for their ability to predict domain-domain and domain-motif interfaces (DDIs and DMIs). We found that prediction accuracies drop when using longer protein fragments or full length proteins for interface predictions and developed a strategy particularly suited for the prediction of novel domain-motif interfaces in human PPIs. We applied this strategy to 62 PPIs from HuRI that connect disease-associated proteins and experimentally assessed the obtained interface predictions for seven PPIs using a plate-based bioluminescence resonance energy transfer (BRET) assay (Trepte et al, 2018) combined with site-directed mutagenesis. We identify novel interface types and report on important limitations and sources of errors in AF-derived structural models, which pave the way for future improvements in the field.

Results

Evaluating AlphaFold's accuracy for predicting domain-motif interfaces

To thoroughly assess the ability of AF to predict structures of binary protein complexes that are formed by a DMI, we extracted information on annotated DMI structures from the ELM DB (Kumar et al, 2022). We selected one representative structure per motif class (136 structures in total), manually defined the minimal domain and motif boundaries, and submitted the corresponding protein sequence fragments for interface prediction to AF (Fig. 1A; Dataset EV1). The domain sequences from this benchmark dataset mostly shared 20–30% sequence identity (Appendix Fig. S1A). To evaluate the accuracy of the predicted structural models, we superimposed the actual structure and predicted model on their domains and based on this superimposition, we computed the all atom RMSD between the motif of the predicted model and the actual structure (Fig. 1A). We found that 35% of the structural models were so accurately predicted that even the side chains of the motif were correctly positioned while for another 32% the backbone but not the side chains of the motif were accurately predicted. For 26% of the structures the motif was modeled into the correct pocket, but in a wrong conformation, while, for the remainder of the structures, AF failed to identify the right pocket (Fig. 1A; Dataset EV1). A similar performance was obtained when using the DockQ metric (Appendix Fig. S1B,C; Dataset EV1). This performance is unaltered when using or switching off AF's template function (Fig. S1D,E). The use of DMI structures annotated by the ELM DB enables us to explore potential differences in AF's performance regarding motif properties. We find no significant differences in average model accuracy between different categories of motif classes (two-sided Mann–Whitney test on all pairwise combinations, n: DEG = 10, DOC = 21, LIG = 94, TRG = 9, MOD = 2, $\alpha = 0.05$, test statistics of all pairwise combinations between 15 and 852, Appendix Fig. S1F), although the variance in model accuracy appears to differ between the motif classes. Similarly, we found no significant difference in prediction accuracy when

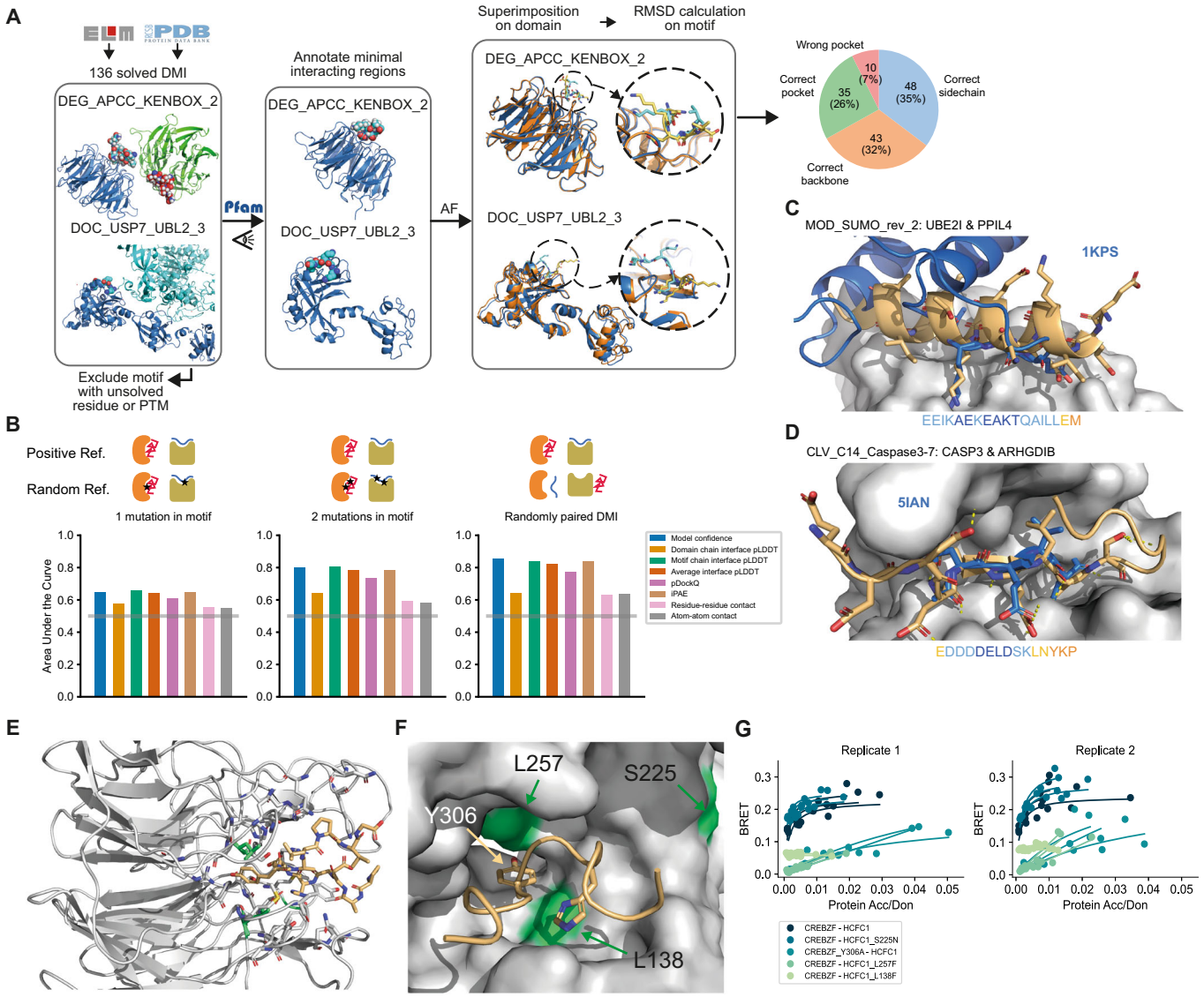


Figure 1. Benchmarking and application of AF for DMI interface prediction using minimal interacting fragments.

(**A**) Schematic illustrating the assembly of the DMI positive reference dataset and evaluation of AF prediction accuracies by superimposition of the solved and modeled structures. Blue and cyan indicate the domain and motif in the native structure, respectively. Orange and yellow indicate the domain and motif in the modeled structure, respectively. Proportion of structures of DMIs predicted by AF to different levels of accuracy is shown on the right. (**B**) Area under the Receiver Operating Characteristics Curve (AUROC) for different metrics using the DMI benchmark dataset as positive reference and the following different random reference sets: Left, 1 mutation introduced in conserved motif position; middle, 2 mutations introduced in conserved motif positions; right, random reshuffling of domain-motif pairs. Gray horizontal line indicates the AUROC of a random predictor. (**C**) Superimposition of AF structural model for motif class MOD_SUMO_rev_2 (orange) with homologous solved structure (PDB:1KPS) from motif class MOD_SUMO_for_1 (blue). The motif sequence used for prediction is indicated at the bottom, colored by pLDDT (dark blue=highest pLDDT). (**D**) Superimposition of AF structural model for motif class CLV_C14_Caspase3-7 (orange) with homologous structure (PDB:5IAN) solved with a peptide-like inhibitor (blue). The motif sequence used for prediction is indicated at the bottom, colored by pLDDT (dark blue=highest pLDDT). (**E**) AF prediction of a LIG_HCF-1_HBM_1 motif in CREBZF (orange) binding to the beta-propeller Kelch domain of HCFC1 (gray). Mutated domain residues for experimental testing are colored in green. (**F**) Close up on the interface shown between CREBZF and HCFC1 from (**E**). Coloring is the same as in (**E**). Key conserved motif residues are drawn as sticks. Mutated residues in the domain and motif for experimental testing are labeled. (**G**) BRET titration curves are shown for wildtype interactions and mutant constructs for CREBZF-HCFC1 pairs for two biological replicates, each with three technical replicates. Protein acceptor over protein donor expression levels are plotted on the *x*-axis determined from fluorescence and luminescence measurements, respectively.

stratifying by the secondary structure elements adopted by the motifs (two-sided Mann–Whitney test on all pairwise combinations, *n*: helix = 42, strand = 7, loop = 87, $\alpha = 0.05$, test statistics of all pairwise combinations between 184 and 2029, Appendix Fig. S1G), nor by how hydrophobic, symmetric, or degenerate the motif sequence is (Pearson r < abs(0.08), $\alpha = 0.05$ Appendix Fig. S1H–J). AF models display significantly more differences to structures solved by other methods, i.e., NMR, than X-ray crystallography (two-sided Mann–Whitney test, *n*: X-ray = 115, Others = 21, $p < 0.01$, test statistics = 811, Appendix Fig. S1K) possibly because

NMR structures better represent structural dynamics that AF cannot capture, since it was trained to predict the crystallized forms of proteins.

The all-atom motif RMSD significantly anti-correlates with various AF-derived metrics (Pearson $r = -0.55$, p-value < 0.05 Appendix Fig. S1L,M; Dataset EV1) suggesting that these metrics are indicative of good versus bad structural models and can be used for de novo interface predictions. To evaluate AF's ability to identify high confident structural models of DMIs, we generated three different random DMI datasets. First, we randomly paired domain and motif sequences from the positive reference dataset taking into account that no motif sequence was paired with a domain sequence from the domain type that the motif is known to interact with. Second and third, we mutated one and two key motif residues, respectively, to residues of opposite chemico-physical properties. Based on the conservation of these key motif residues, we assume that the mutations would be disruptive to binding, at least when experimentally tested using minimal interacting protein fragments. Receiver operating characteristic (ROC) and precision-recall (PR) curves using the positive and random datasets (Fig. 1B; Appendix Fig. S2A,B; Dataset EV2) show that the domain interface residue pLDDT (for all metric definitions, see Methods) or the number of atoms or residues predicted to be in contact with each other, discriminated poorly between all reference datasets (AUC around 0.64). Furthermore, we observed that all tested metrics failed to discriminate interacting from non-interacting interfaces when mutating one motif residue (max AUC 0.66). However, the AF-derived metrics model confidence (preprint:Evans et al, 2021), average interface residue pLDDT, average motif interface residue pLDDT, pDockQ (Bryant et al, 2022), and iPAE (Teufel et al, 2023) discriminated well between both reference datasets when randomizing domain-motif pairs or introducing two motif mutations (max AUC 0.86, ROC statistics and ideal cutoffs can be found in Dataset EV2). We also evaluated whether the top 5 reported models by AF tend to be more similar to each other when corresponding to a correct structural model (Pozzati et al, 2022) and found that this feature has moderate predictive power (Appendix Fig. S2C).

Application of AlphaFold for providing structural models for motif classes without available structural data

After evaluating the accuracy of AF to predict DMIs using minimal interacting regions, we aimed to use this setup for the prediction of structural models for motif classes in the ELM DB for which no structure of a complex has been solved yet. We identified 125 such motif classes based on ELM DB annotations. Of those, we selected all domain-motif instances where both the motif and the domain were derived from human or mouse proteins and submitted the corresponding domain and motif sequences for structure prediction to AF (Dataset EV3). Using a motif chain pLDDT cutoff of > 70, we obtained confident structural models for 21 motif classes. We manually inspected the structural models and noticed that even though these ELM classes have no annotations with structures, solved structures for an exact ELM instance or a very likely new instance for the ELM class are available for 11 out of the 21 cases. For most others, a close homolog structure had been solved, i.e., for LIG_MYND_3 and LIG_MYND_1, a structure solved by NMR for a LIG_MYND_2 interaction is available (Appendix Fig. S2D,E). For MOD_SUMO_rev_2, a structure of a reversed motif is available (and annotated as such in the MOD_SUMO_for_1 class). Here it is interesting to see how very dissimilar binding modes (flexible for MOD_SUMO_for_1, helical for MOD_SUMO_rev_2), are still able to place the important binding residues in the same pockets (Fig. 1C). For CLV_C14_Caspase3-7, the structure of the caspase bound to peptide-like inhibitors has been solved (e.g. PDB:1F1J, PDB:5IAN, PDB:6KMZ), and structures of more distant caspases bound to a cleaved peptide substrate are also available. For proteases, one great advantage of AF is the ability to model both the catalytically active enzyme and an uncleaved substrate, which is practically impossible to solve experimentally (Fig. 1D).

Finally, for LIG_HCF-1_HBM_1 we were not able to identify a homologous structure in the PDB, hence, our AF-derived structural models for this motif class are likely novel. Motifs of this class are bound by the N-terminal beta-propeller Kelch domain of HCFC1 consisting of six Kelch repeats. Kelch domains have been shown to bind to motifs at a number of different sites, and thus, without prior knowledge, it is difficult to determine where the HCFC1-binding motif (HBM) would bind. HCFC1 is a transcription factor that associates with other transcription factors (Lu et al, 1997), splice factors (Ajuh et al, 2002), and cell cycle regulators (Freiman and Herr, 1997; Machida et al, 2009). We generated AF models of high confidence for the HCFC1 Kelch domain interacting with multiple motif instances that are annotated in the ELM DB. All complexes show the tyrosine of the motif docked into a deep pocket at the bottom/top of the Kelch domain (Fig. 1E,F; Appendix Fig. S2F–H), with slight variations in how the tyrosine is exactly positioned in the pocket (Fig. S2F–H). Based on clone availability we selected the structural model between HCFC1 and CREBZF for experimental validation. For this purpose, we used a BRET protein interaction assay that is based on transient overexpression of two proteins in HEK293 cells (Trepte et al, 2018). Both proteins are expressed as fusion constructs either to the Nanoluc luciferase (the donor) or mCitrine (the acceptor). Interaction of both proteins results in a BRET from the oxidized substrate of the donor to the acceptor molecule, if both are close enough to each other for the BRET to occur (see Methods for details). We observed significant binding and BRET saturation when assaying wildtype CREBZF and HCFC1 proteins (Fig. 1G; Appendix Fig. S2I,J). Mutation of the [DE]H.Y motif tyrosine to alanine (Y306A) or mutation of two residues in the Kelch domain pocket (L257F, L138F), which are modeled to be in contact with the motif tyrosine or histidine residue (Fig. 1F), strongly reduced BRET signals indicating weakening or loss of binding (Fig. 1G; Appendix Fig. S2I,J). A pathogenic mutation (S225N, source ClinVar (Henrie et al, 2018)) close to the pocket slightly reduced expression levels of HCFC1 but did not result in loss of binding (Fig. 1F,G; Appendix Fig. S2I,J). Our experiments suggest that a potential pathogenic mechanism of this mutation is not mediated via perturbed binding of partners to the Kelch repeat domain pocket of HCFC1 that we identified in this study. Unfortunately, no assertion criteria for the annotation of this mutation to be pathogenic is provided by ClinVar meaning that the mutation is either not pathogenic after all or its pathogenicity is mediated via another perturbed function not tested in this study. Collectively, these experimental results support the structural models of the HCFC1 Kelch domain pocket - motif interaction and overall provide highly confident structural models for multiple motif classes of the ELM DB without available structural information (Dataset EV4).

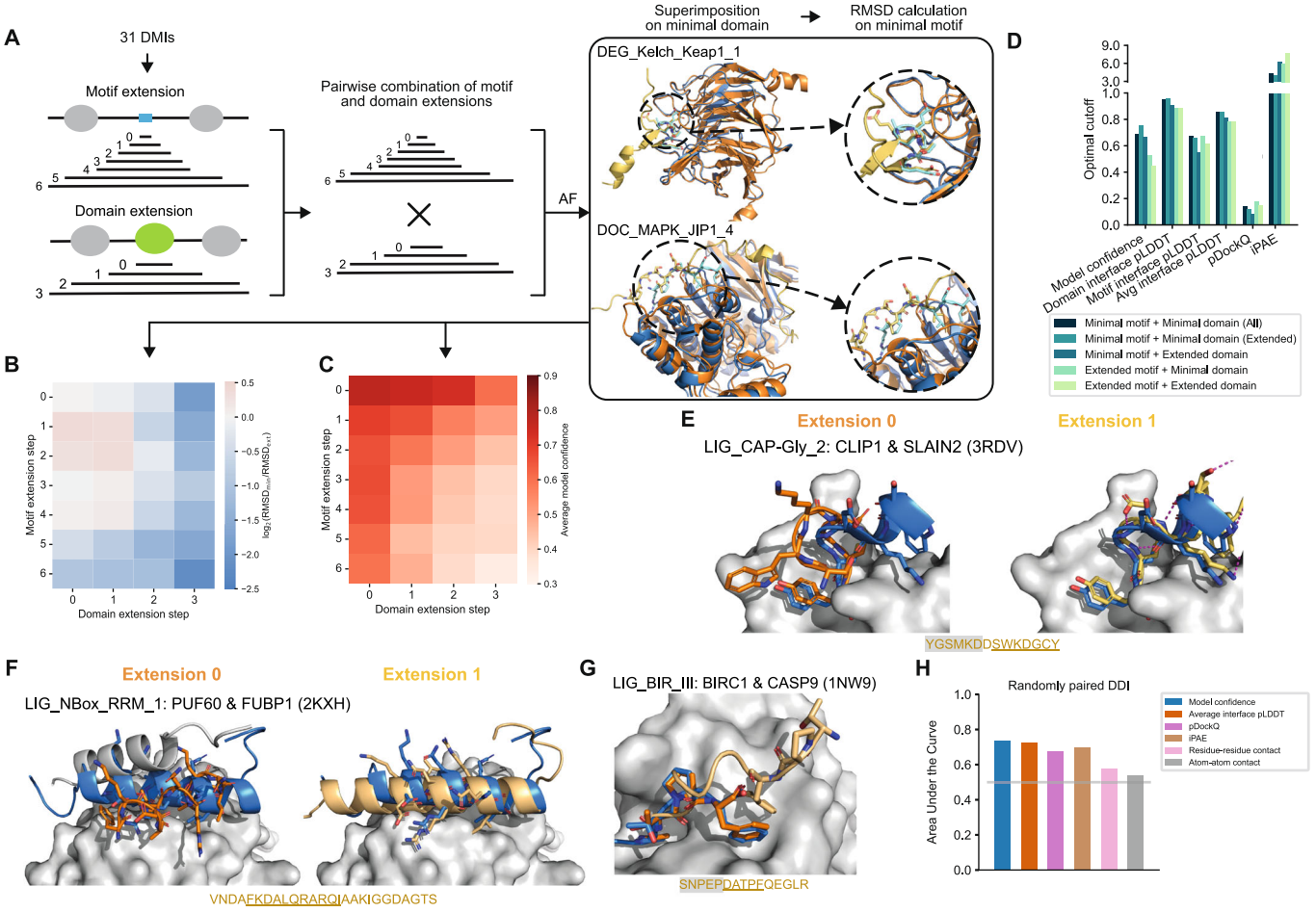


Figure 2. Effect of protein fragment extensions on the accuracy of AF predictions.

(A) Workflow established to assess changes in AF performance upon protein fragment extension. Blue and cyan indicate the domain and motif in the native structure, respectively. Orange and yellow indicate the domain and motif in the modeled structure, respectively. **(B)** Heatmap showing the fold change in motif RMSD before and after extension where positive values indicate improved predictions from extension and negative values indicate worse prediction outcomes upon extension. **(C)** Heatmap of the average model confidence for combinations of different motif and domain sequence extensions. **(D)** Optimal cutoffs derived for different metrics from ROC analysis benchmarking AF different motif and domain extensions from the reference dataset used in A and random pairings of domain and motif sequences. pLDDT-related metrics were divided by 100 for visualization purposes. **(E, F)** Superimposition of the structural model of the minimal (left, orange) or extended (right, yellow) motif sequence with the solved structure (motif in blue) for two different motif classes as indicated on the top of each panel. The motif sequence from the solved structure is indicated at the bottom. Motif residues are underlined, motif residues not resolved in the structure have a gray background. Sticks indicate the motif residues, domain surfaces are shown in gray based on experimental structures. **(G)** Superimposition of the structural model of the minimal (orange) and extended (yellow) motif sequence with the solved structure (motif in blue) for a motif instance from the motif class LIG_BIR_III. Motif sequence indicated as in **(E)**. **(H)** Area under the Receiver Operating Characteristics Curve (AUROC) for different metrics using the DDI benchmark dataset as positive reference and randomly shuffled domain-domain pairs as random reference. Gray horizontal line indicates the AUROC of a random predictor.

Evaluation of AlphaFold's ability to predict interfaces in full length proteins

Most PPIs known to date have been identified using full length protein sequences in systematic interactome mapping efforts. For the vast majority of these PPIs, no fragment or interface information is available. Thus, the question emerges how AF would perform on DMI predictions when longer protein sequences or full length proteins are submitted. To answer this question we selected 31 DMI structures from the positive reference dataset used above and generated random domain-motif pairs of those as negative control. The selected structures were sampled from different prediction accuracy categories (Fig. 1A; Dataset EV5). We then gradually extended the motif and domain sequences by first adding flanking disordered regions, then neighboring folded domains before using the full length sequences (Fig. 2A). Comparison of the motif RMSD computed for extended versus minimal domain-motif pairs from the positive reference dataset revealed that the addition of flanking disordered regions on the motif or domain side sometimes slightly improved prediction accuracies while the addition of neighboring structured domains or the use of full length sequences led to a significant worsening of model accuracies (Fig. 2B; Dataset EV5). Interestingly, despite the fact that, for smaller extensions, model accuracies remained the same or slightly improved as determined by motif RMSD, AF-derived metrics such as the model confidence or average motif

interface residue pLDDT gradually dropped with increasing fragment length (Fig. 2C; Appendix Fig. S3A-C). ROC plots of predictions for a benchmark consisting of the positive and random domain-motif pairs revealed that, upon extension, the optimal cutoff of model confidence and iPAE considerably changed as well (Fig. 2D; Appendix Figs. S3D,E, S4A; Dataset EV6). This means that different model confidence or iPAE cutoffs are to be used depending on the length of the submitted protein sequences, which is rather impractical and thus disfavors both metrics for DMI predictions. The average motif interface residue pLDDT metric appeared to be more robust with respect to fragment length. Based on these results we chose this as the main metric and a cutoff of 70 to discriminate good from bad AF-generated DMI models regardless of fragment length.

Extending motif sequences for interface prediction with AlphaFold reveals important motif sequence context

Various studies have highlighted that flanking sequences of motifs can influence binding affinities and specificities (Luck et al, 2012; Bugge et al, 2020). Motif annotations in the ELM DB usually refer to the core sequence of the motif, often because information on putative roles of flanking sequences is missing. In the previous section, we observed that some motif extensions notably improved AF prediction accuracies. In the hope that these cases would point to motifs with important sequence context, we manually inspected eight predictions for which the motif RMSD decreased by more than 1 Å when extending the minimal motif sequence once to the left and right by the length of the motif (extension step 1 in Fig. 2A; Appendix Fig. S4B).

By doing so interesting patterns emerged: The most prevalent contribution to increased prediction accuracies is the stabilization of the secondary structure of the motif contributed by both sidechain and backbone atoms in the flanking regions, as shown for the interaction involving the motif LIG_CAP-Gly_2 (Fig. 2E; Appendix Fig. S4C). For the LIG_NBox_RRM_1 motif, AF placed a part of the domain into the binding pocket rather than the motif, although the motif had the correct helical conformation. Elongation of the motif extended this helix, thereby increasing the interaction surface and eventually pushing out the domain's tail from the pocket (Fig. 2F). This fits with other reports where AF has been shown to predict preferential binding of competing motifs (Chang and Perez, 2023). For the LIG_HOMEOBOX class prediction, the motif is positioned in the wrong pocket unless flanking regions are included (Appendix Fig. S4C). For DOC_MAPK_JIP1_4, motif extension results in an extended motif conformation and consequently in a structural model with lower overall RMSD (Appendix Fig. S4C). For the LIG_GYF class, most models converge into an inverse orientation of the backbone except for one of the extended motifs, which lies in the binding pocket in the correct orientation (Appendix Fig. S4C). In summary, these analyses point to motif classes whose sequence boundaries could be refined.

Interestingly, for a motif instance from the LIG_BIR_III_2 class, slight motif extensions actually led to a substantial decrease in prediction accuracy. In this case, the motif is located at a neo-N-terminus that is only revealed after cleavage of the protein by a caspase (Fig. 2G). When the motif is extended in the context of the full length protein, the residues now upstream of the previous neo-N-terminus likely impede binding of the motif into the pocket due to steric clashes. AF predicts the extended motif to bind in reversed orientation and it is mostly pushed out of the pocket. This highlights the importance of not only incorporating sequence context but also knowledge about the biological context, wherever possible, into AF modeling and model interpretation.

Evaluating AlphaFold's performance for the prediction of domain-domain interfaces

Folded domains can not only interact with motifs but also with other folded domains forming so-called domain-domain interfaces (DDIs). To enable simultaneous prediction of DDIs and DMIs in a given protein interaction, we set out to evaluate AlphaFold's performance on DDI predictions using a reference dataset of 48 DDI structures that we manually curated out of random selections of domain-domain contact pairs extracted from 3did (Mosca et al, 2014). As a negative dataset, we randomized the pairing of these domains. Using ROC and PR statistics we found that AlphaFold performed slightly worse on this DDI benchmark dataset compared to its performance on DMIs (max AUC 0.73 vs. 0.86) (Fig. 2H; Appendix Fig. S4D–F; Dataset EV7) but still showed significant discriminative power. Interestingly, the best performing metric for DDI predictions was the average interface pLDDT score with an optimal cutoff of 75, which ranked fourth for DMI predictions.

Comparison of AlphaFold v2.2 with v2.3

During the course of our work, AF multimer version 2.3 was released. To determine whether the new release improved DMI and DDI prediction accuracies, we repeated all benchmarking with AF v2.3 and found that motif RMSDs and other AF-derived metrics on average improved compared to AF v2.2 when using minimal interacting fragments (Appendix Fig. S5A–D; Dataset EV1, two-sided Wilcoxon signed-rank test on motif all atom RMSD: $n = 136$, $W = 2413$, $p < 0.0001$). AF v2.3 still showed a decrease in prediction accuracy when using extended protein fragments but this decrease was less pronounced compared to the corresponding decrease for v2.2 (Appendix Fig. S5E,F; Dataset EV5). Despite these improvements on the sensitivity side of AF, when benchmarked against random datasets, overall prediction accuracies only slightly improved compared to v2.2 (Appendix Fig. S5G,H; Appendix Fig. S6A–C; Dataset EV2, EV6, EV7, EV8).

Application of AlphaFold for the discovery of novel interfaces in protein interactions without any a priori interface information

Since the use of larger or full length protein sequences leads to a poor sensitivity for DMI predictions by AF, we devised the following strategy for the use of AF for interface predictions for known protein interactions: Using AF models of the full length monomeric structures of both interacting proteins, we decided on boundaries between structured domains and disordered regions based on manual inspection (see Methods). We then fragmented the disordered regions by designing overlapping fragments varying in length from ten residues up to the length of the respective disordered region (Fig. 3A). We then paired disordered with ordered, and ordered with ordered fragments for interface prediction by AF (Fig. 3A). To assess to which extent this

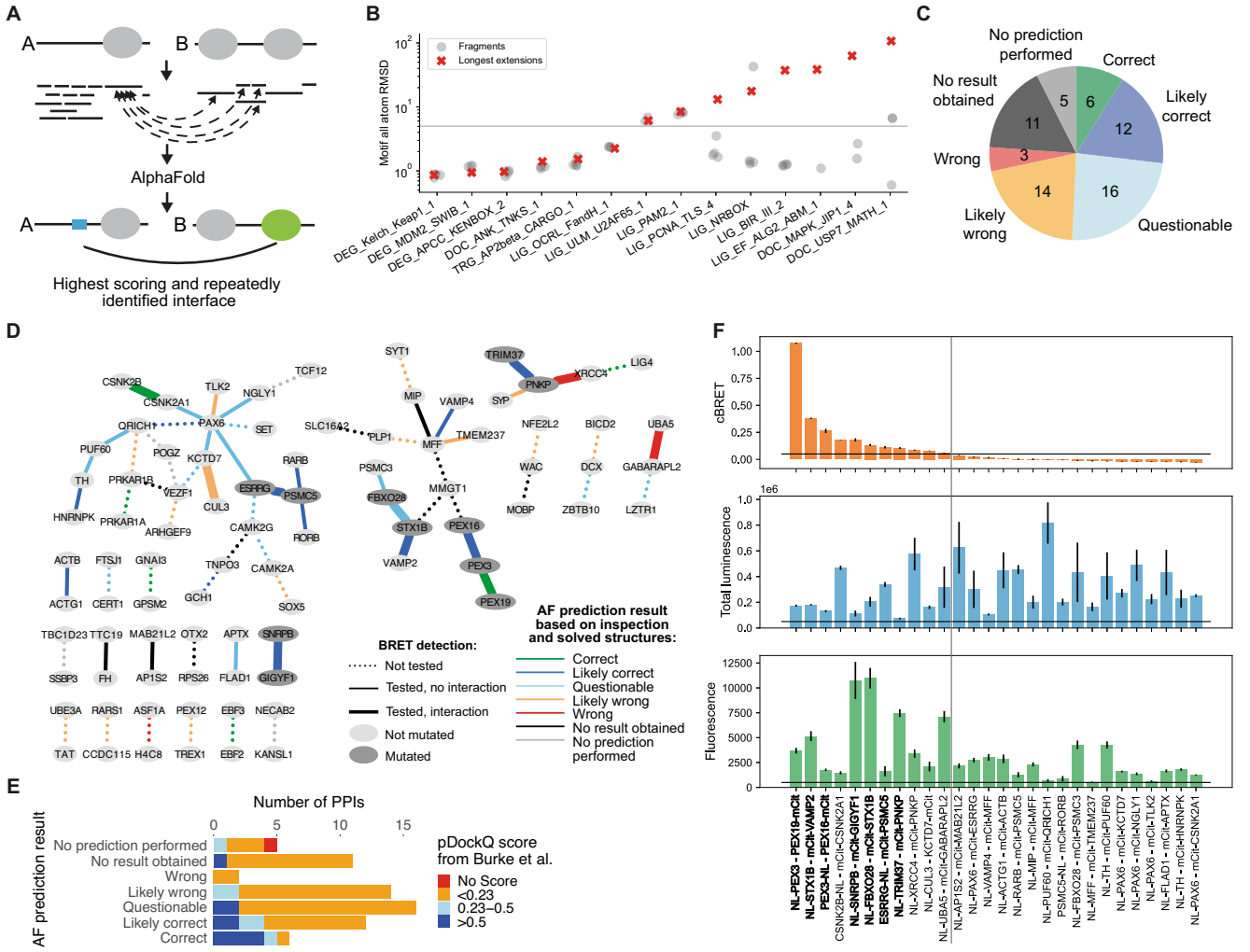


Figure 3. AF prediction and experiments on PPIs connecting NDD proteins.

(**A**) Schematic of the fragmentation approach applied on a pair of interacting proteins, A and B. Proteins are fragmented into folded and disordered regions based on manual inspection. Disordered regions are further fragmented. All disordered and folded fragments of one protein are paired with the folded regions of the other protein and vice versa for AF prediction. (**B**) Accuracy measured in motif RMSD compared to native structures for models obtained from fragmenting proteins from 20 DMIs from the positive reference dataset and comparison to model accuracy obtained when using (near) full length proteins for structure prediction (red crosses). Only models that meet the cutoff for identifying high confident models are shown. Six DMIs did not result in any such model. The gray horizontal line indicates the RMSD cutoff used to identify accurate models (see methods for details). (**C**) AF prediction outcome on 67 HuRI PPIs connecting NDD proteins. (**D**) PPI networks illustrating AF prediction outcomes and experimental retesting of PPIs in BRET assay. (**E**) Number of PPIs connecting NDD proteins with structural models at indicated pDockQ cutoffs from (Burke et al, 2023) grouped based on AF prediction outcomes using the fragmentation approach as shown in (**C**). (**F**) cBRET, total luminescence, and fluorescence for 28 PPIs connecting NDD proteins that were tested in the BRET assay. Luminescence and fluorescence measurements indicate expression levels of NL and mCit fusion proteins, respectively. Black horizontal lines indicate expression level and PPI detection cutoffs. The gray vertical line separates the detected (left) from undetected PPIs. Protein pairs in bold indicate those selected for interface validation via site-directed mutagenesis. Error bars indicate STD of three technical replicates. Source data are available online for this figure.

fragmentation approach would lead to an increase in sensitivity but also in false model predictions, we selected 20 out of the 31 DMI structures that were previously used to investigate the effect of fragment extension on prediction accuracies. We attempted model prediction with the full length sequences of these 20 DMI pairs and obtained a model for two of which only one met the motif interface pLDDT cutoff and corresponded to an accurate prediction (TRG_AP2beta_CARGO_1 in Fig. 3B; Dataset EV9, see methods for details). We then switched to using fragment extension step 5 for motifs and/or 2 for domains (Fig. 2A) and obtained accurate models for an additional 5 of the 20 DMI pairs. Applying the full fragmentation approach onto all 20 DMI pairs resulted in accurate model prediction for an additional 6 DMI pairs (Fig. 3B) representing an increase in sensitivity for full length vs fragments from 5 to 60%. We then shuffled the 20 DMI pairs to generate 20 random DMI pairs for which we performed the fragmentation approach. As expected from an earlier estimated 20% false positive rate (FPR) (Appendix Fig. S4A), 19 of the 20 random protein pairs had at least one fragment pair that produced a model above the motif interface pLDDT cutoff (Appendix Fig. S6D; Dataset EV9)

indicating that predictions done using this fragmentation approach can substantially increase sensitivity while also producing a considerable number of false models using the established scoring metrics. This needs to be taken into account when modeling new interactions with this fragmentation strategy, as covered in the following section.

We selected PPIs from HuRI that connect proteins associated with neurodevelopmental disorders (NDDs) and subjected these to our AF fragmentation pipeline to predict putative DMIs and DDIs. For 51 out of 62 PPIs we obtained at least one structural model of significant confidence (Fig. 3C,D). In retrospect, manual inspection of the predictions obtained for these PPIs revealed that, for 9 PPIs, a solved structure of the interface was already available. Reassuringly, six out of these were accurately predicted by AF. For the remainder of the PPIs, 12, 16, and 14 resulted in a likely correct, questionable, or likely wrong prediction, respectively, based on manual inspection of the models (Fig. 3C,D; Dataset EV10). Likely wrong predictions were scored as such based on docking of the protein partner into nucleic acid or metal ion binding or catalytically active sites. We also considered structural models as likely wrong, if different protein fragments of the partner were predicted with similarly high scores to bind to the same pocket on the domain. More detailed information can be found in Methods and Appendix Text S1. Of note, for 8 of the 12 PPIs with a likely correct prediction, AF predictions performed using the full length proteins (Burke et al, 2023) did not result in a high confidence prediction (Fig. 3E). 28 of the 62 PPIs were in our hands amenable to experimental testing using the BRET assay introduced earlier (see Methods for details). Significant BRET signals were observed for 11 of these 28 PPIs (Fig. 3F). Of those, 7 PPIs were selected for validating the predicted interfaces (Fig. 3D,F). The remaining four PPIs were not further considered because for three of them a structure already exists (CSNK2B-CSNK2A1, PNKP-XRCC4, UBA5-GABRAPL2) and for the fourth interaction (KCTD7-CUL3) we classified the predicted interface as likely wrong. Next, we will first describe failures in validating predicted interfaces followed by the successes.

For the interaction between PNKP and TRIM37, we obtained high confident structural models involving two different interfaces. AF predicted the PNKP FHA domain to bind to several disordered stretches in TRIM37 (Fig. 4A) that are overall negatively charged. These short regions were predicted to bind to a pocket on the FHA domain that is known to bind phosphorylated threonines (Durocher et al, 2000), which led us to conclude that these predictions were likely wrong. AF also predicted the MATH domain of TRIM37 to bind to two separate disordered putative motifs located between the FHA domain and phosphatase domain in PNKP (Fig. 4A–C). However, none of the mutants aimed at disrupting the predicted interfaces (Fig. 4B) involving the MATH domain showed a decrease in BRET signal compared to wildtype (Fig. 4D; Appendix Fig. S7A) indicating that TRIM37 and PNKP do not interact with each other via this interface.

AF predicted with high confidence binding of PSMC5 to the hormone receptor domain of ESRRG via two distinct motifs (Fig. 4E–G) with similarity to LxxLL motifs known to bind this type of domain (LIG_NRBOX in ELM DB). We reproducibly found that none of the motif mutations in PSMC5 decreased binding to ESRRG compared to wildtype while both domain pocket mutations led to a remarkable reduction in BRET signal (Fig. 4H; Appendix Fig. S7B,C) indicating that PSMC5 might bind to ESRRG via this pocket but not with the predicted motifs.

AF predicted a coiled-coil interface between STX1B and VAMP2 of moderate confidence (Fig. 5A,B). STX1B is a close homolog to STX1A, which binds in a 4-helix bundle to VAMP2 together with SNAP25 in a 1:1:2 stoichiometry, respectively, as observed by crystallography (PDB:1N7S (Ernst and Brunger, 2003)). This structure together with our predictions suggest that STX1B might bind VAMP2 in a similar way. Indeed, removal of the single helical SNARE domain in STX1B led to complete loss of binding to VAMP2 (Fig. 5C; Appendix Fig. S8A,B). Interestingly, FBXO28 was predicted by AF to bind to STX1B via a similar coiled-coil interface involving an extended helix in FBXO28 and the SNARE domain in STX1B (Fig. 5A,D). Here, deletion of the SNARE domain in STX1B or of the extended helix in FBXO28 reproducibly reduced, but did not abolish the interaction between STX1B and FBXO28 (Fig. 5E; Appendix Fig. S8C,D). We identified three pathogenic or likely pathogenic mutations in the SNARE domain of STX1B in ClinVar of which V216E and G226R are associated with generalized epilepsy with febrile seizures plus, type 9. Testing all three mutations in the BRET assay we observed a drastic decrease in binding for STX1B V216E to FBXO28 (Fig. 5F; Appendix Fig. S8C,D). However, the measured effects of the mutations on the FBXO28-STX1B interaction do not correlate with their location at the predicted interface. V216E, for example, is not predicted to be in contact with residues of FBXO28 (Fig. 5D). This indicates that the actual predicted orientation of the two extended helices with respect to each other is likely incorrect.

The fact that the deletion of the extended helix in FBXO28 or the SNARE domain in STX1B reduced but did not abrogate binding of both proteins to each other (Fig. 5E) suggests that a secondary interface might exist. Indeed, AF predicted additional interfaces between FBXO28 and STX1B involving folded and disordered regions in both proteins (interfaces i and ii in Fig. 5A). Mutations designed to disrupt these interfaces partially confirmed the involvement of some of these regions in binding as assayed with BRET (Appendix Fig. S8E–H). In addition, the pathogenic mutation R348L in FBXO28 predicted to be at interface ii seemed to increase binding to STX1B (Appendix Fig. S8I–L). In summary, our experimental data indicate that multiple regions of FBXO28 and STX1B may be involved in the binding but the exact structural details of this interaction remain to be elucidated. In the following two sections, we will describe in more detail successful interface validations for interactions involving PEX3, PEX19, and PEX16 as well as SNRPB and GIGYF1.

PEX3, PEX19, and PEX16

The interaction interface between PEX19 and PEX3 has been structurally resolved before and consists of an interaction between an N-terminal motif in PEX19 that binds to the cytosolic alpha-helical domain of PEX3 (PDB:3MK4, (Schmidt et al, 2010)). Using corresponding protein fragments, AF predicted a structural model that is highly similar to the solved structure (Fig. 5G; Appendix Fig. S9A,B). We introduced mutations in the PEX19 motif and PEX3 pocket (Appendix Fig. S9A) and found that F29K in the motif weakened but clearly maintained BRET binding signals indicating the existence of a secondary binding site between both proteins (Fig. 5H; Appendix Fig. S9C,D). Indeed, AF predictions with other

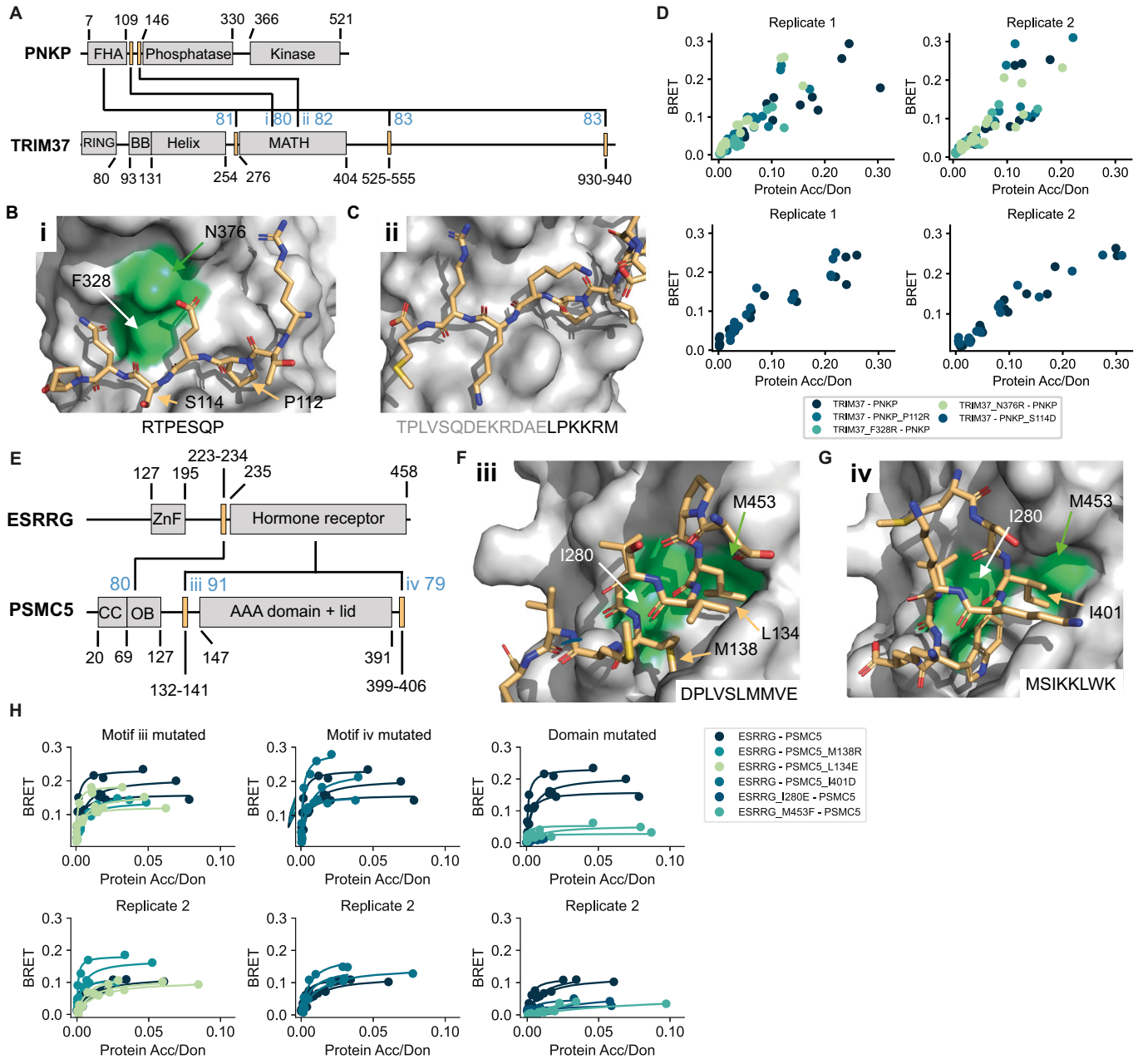
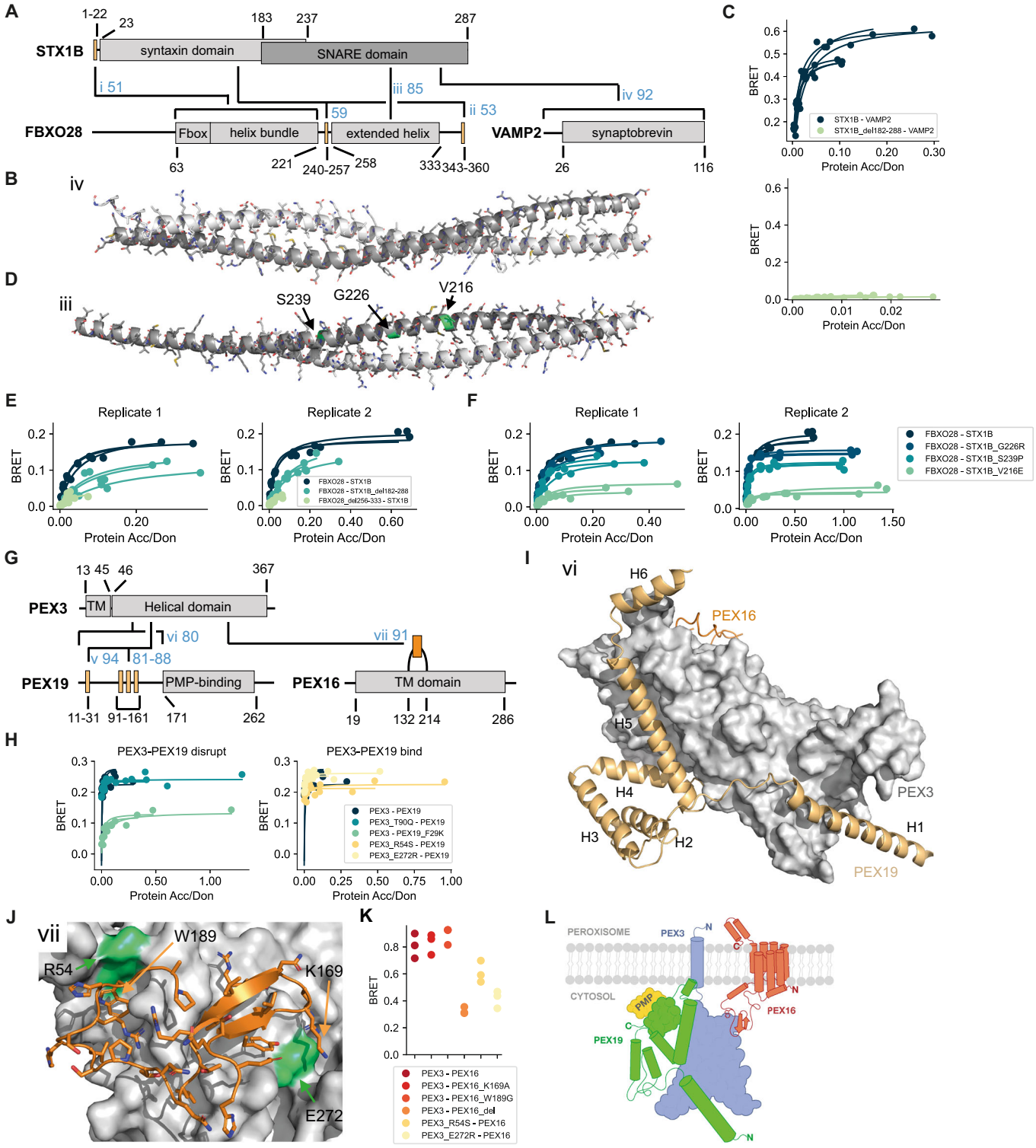


Figure 4. Verification of interface predictions for TRIM37-PNKP and ESRRG-PSMC5.

(A) Schematic of the domain architecture of PNKP and TRIM37 with indication of top predicted interfaces. Numbers in blue indicate the motif interface pLDDT for the respective interface. Roman numbering refers to structural models in **(B)** and **(C)**. **(B)** Structural model of interface i shown in **(A)** with labeled residues that were mutated. **(C)** Structural model of interface ii shown in **(A)**. **(D)** BRET titration curves are shown for wildtype interaction and mutants for two biological replicates, each with three technical replicates. Protein acceptor over protein donor expression levels are plotted on the *x*-axis determined from fluorescence and luminescence measurements, respectively. The BRET trajectory could not be fitted because of an unusual saturation behavior (see methods for details). **(E)** Schematic of the domain architecture of ESRRG and PSMC5 with indication of top predicted interfaces. Numbers in blue indicate the motif interface pLDDT for the respective interface. Roman numbering refers to structural models in **(F)** and **(G)**. **(F)** Structural model of interface iii shown in **(E)** with labeled residues that were mutated. **(G)** Structural model of interface iv shown in **(E)**. **(H)** BRET titration curves are shown for wildtype interaction and mutants of ESRRG-PSMC5 pairs for two biological replicates, each with three technical replicates. Protein acceptor over protein donor expression levels are plotted on the *x*-axis determined from fluorescence and luminescence measurements, respectively. In panels **(B)**, **(C)**, **(F)**, and **(G)** motif sequences are indicated at the bottom. Gray letters indicate residues not predicted to bind. Source data are available online for this figure.



disordered fragments of PEX19 paired with the PEX3 domain resulted in highly confident models for interfaces involving a binding pocket on PEX3 that is distal to the pocket where the N-terminal PEX19 motif is known to bind. When using a protein fragment that spans the full disordered N-terminal region of PEX19 (1–170), AF predicts the known PEX3-binding motif and helix 4 and 5 to dock into the primary and secondary pocket, respectively (Fig. 5G,I), supporting simultaneous interaction via both interfaces.

While the interaction between PEX3 and PEX16 has been described before, little is known about how both proteins interact with each other. The monomeric AF model of PEX16 shows a helical fold, which could in its entirety be transmembrane (TM).

Figure 5. Verification of interface predictions for STX1B-FBXO28, STX1B-VAMP2, PEX3-PEX19, and PEX3-PEX16.

(A) Schematic of the domain architecture of STX1B, FBXO28, and VAMP2 with indication of top predicted interfaces. Numbers in blue indicate the motif interface pLDDT (for order-disorder fragment pairs) or average interface pLDDT (for ordered-ordered fragment pairs) for the respective interface. Roman numbering refers to structural models in **(B)**, **(D)**, Appendix Fig. S8E, and Appendix Fig. S8I. **(B)** Structural model of interface iv shown in **(A)**. In panel **(B)** and **(D)**, the chains are color-coded according to the colors of the domains in **(A)**. **(C)** BRET titration curves are shown for wildtype interactions and deletion constructs for two biological replicates, each with three technical replicates. Protein acceptor over protein donor expression levels are plotted on the x-axis determined from fluorescence and luminescence measurements, respectively. **(D)** Structural model of interface iii shown in **(A)** with tested pathogenic mutations labeled and colored in green. **(E, F)** BRET titration curves are shown for wildtype interactions and deletion constructs for two biological replicates, each with three technical replicates. Protein acceptor over protein donor expression levels are plotted on the x-axis determined from fluorescence and luminescence measurements, respectively. **(G)** Schematic of the domain architecture of PEX3, PEX19, and PEX16 with indication of top predicted interfaces. Numbers in blue indicate the motif interface pLDDT for the respective interface. Roman numbering refers to structural models in **(I)**, **(J)**, and Appendix Fig. S9A. Region vi covers residues 1–170, which includes the previously reported N-terminal motif as well as three putative motifs suggested by the AF models. **(H)** BRET titration curves are shown for wildtype interaction and mutants of PEX3-PEX19 pairs for three technical replicates. Protein acceptor over protein donor expression levels are plotted on the *x*-axis determined from fluorescence and luminescence measurements, respectively. The left plot displays mutants aimed at disrupting binding between PEX3-PEX19 while the right plot displays mutants aimed at disrupting the PEX3-PEX16 PPI why binding between PEX3-PEX19 should not be altered. **(I)** Superimposition of structural models of interface vi (PEX3-PEX19) and vii (PEX3-PEX16) on the PEX3 domain. Note that modeling smaller fragments of PEX19 generates alternative interactions with the binding sites. **(J)** Structural model of interface vii shown in **(G)**. **(K)** BRET values with subtracted bleedthrough for PEX3-PEX16 wildtype and various mutated constructs. Three technical replicates are shown. **(L)** Proposed model for how the trimeric complex of PEX3, PEX19, and PEX16 might assemble at the peroxisomal membrane. Source data are available online for this figure.

Between the putative TM helix 4 and 5 there is a large loop (132–214), which was predicted by AF with very high confidence to bind to a third pocket on the PEX3 domain, opposite to both binding sites mentioned earlier for PEX19 (Fig. 5G,I,J). Of note, different fragments of this loop as well as the entire PEX16 were repeatedly predicted to bind in similar modes to PEX3, further increasing the confidence in this prediction. Encouraged by these results, we submitted all three full length PEX sequences for complex prediction to AF and obtained a model that supports simultaneous binding of PEX16 and PEX19 to PEX3 (Appendix Fig. S9E). We individually mutated two residues in the PEX16 loop, deleted the loop in its entirety (del162-192), and mutated two residues on PEX3 (highlighted in Fig. 5J). Unfortunately, higher expression levels of PEX16 seem to trigger degradation of PEX3 (Appendix Fig. S9F), which we did not observe for the same constructs when co-expressed with PEX19 (Appendix Fig. S9G). As a consequence, we could not obtain titration curves and BRET50 estimates but obtained reliable BRET signals for lower PEX3-PEX16 DNA transfection ratios showing that the deletion as well as both PEX3 mutants significantly decreased binding to PEX16 (Fig. 5K; Appendix Fig. S9H). Of note, these PEX3 mutants (R54S and E272R) did not alter binding to PEX19, showing that the overall structural integrity of PEX3 was not perturbed by these mutations (Fig. 5H; Appendix Fig. S9D).

PEX3 and PEX19 are peroxin proteins that regulate peroxisome homeostasis. PEX16 is believed to serve as an integral membrane-bound receptor for PEX3 (Matsuzaki and Fujiki, 2008) while PEX3 is thought to serve as a docking site for PEX19 (Fujiki et al, 2006). PEX19 in turn is a cytosolic carrier for peroxisomal membrane proteins to the peroxisome (Fujiki et al, 2006). Combining results from previously published functional studies with the structural and experimental results obtained in this study, a model for a trimeric complex between PEX3, PEX19, and PEX16 emerges (Fig. 5L) where PEX16 fully inserts into the peroxisome membrane via a fold that consists of seven helices (residues 19-286) with its N-terminal end being cytosolic and its C-terminal end protruding into the peroxisome. The extended loop between TM helix 4 and 5 reaches into the cytosol and docks onto PEX3, which is further anchored into the peroxisomal membrane via its N-terminal TM helix (residues 13–45). PEX19 docks onto PEX3, opposite to where PEX16 is bound, via two interaction surfaces—one corresponding to the known PEX3-binding motif in PEX19 and a second one corresponding to a novel motif (residues 99–146) docking at a hitherto unknown second binding site on PEX3 for PEX19. This model explains how PEX3 is anchored to the peroxisomal membrane via PEX16 and how PEX3 can bind very tightly PEX19, which can then deliver PMPs to the peroxisome. Mutations in any of the three PEX proteins are associated with severe developmental phenotypes referred to as peroxisome biogenesis disorders (Fujiki et al, 2022). The vast majority of the around 150 mutations annotated for the three proteins are uncharacterized (Henrie et al, 2018), dozens of which fall into the predicted interfaces. The structural models obtained from this work can inform future studies aimed at characterizing the effects of these mutations.

SNRPB and GIGYF1

AF predicted two different types of interfaces with high confidence for the interaction between SNRPB and GIGYF1. The first interface involves the LSM domain of SNRPB which was predicted to bind to various fragments in the long disordered regions of GIGYF1 (Fig. 6A). These regions do not display any common sequence pattern. The structure of SNRPB has been resolved as part of the Sm ring complex that binds small nuclear RNA (PDB:4WZJ, (Leung et al, 2011)) showing that the surface on the LSM domain predicted to bind to disordered fragments of GIGYF1, is actually engaged in binding LSM domains of other Sm proteins within the complex (Fig. 6B). We thus conclude that these predictions are likely wrong. The second type of interface predicted by AF involves the GYF domain in GIGYF1 and multiple short disordered fragments in the C-terminal region of SNRPB, which repeatedly carry the sequence PPPGM(R) (Fig. 6A,C). We designed various deletion constructs of SNRPB that would gradually remove more and more of the repeated proline-rich motif. We observed, using the BRET assay, that these deletion constructs gradually decreased binding to GIGYF1 (Fig. 6D; Appendix Fig. S10A,B). We also mutated the GYF domain pocket and found that W498E but not L508F would decrease binding to SNRPB (Fig. 6D,E; Appendix Fig. S10A–D). To further corroborate these findings we performed a co-immunoprecipitation experiment, where endogenous GIGYF1 interacted with HA-tagged full length SNRPB (Fig. 6F). This

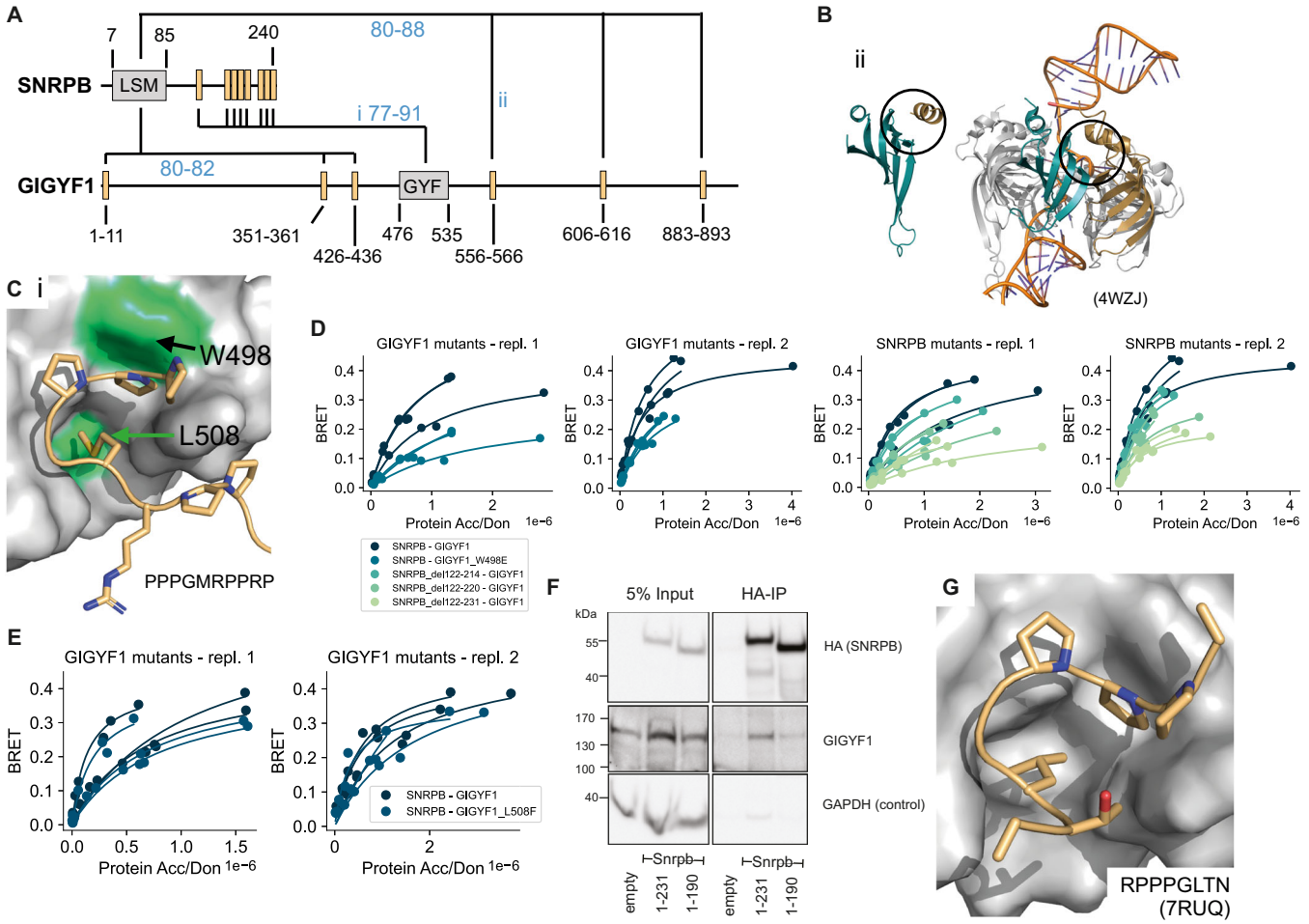


Figure 6. Verification of interface predictions for SNRPB-GIGYF1.

(A) Schematic of the domain architecture of SNRPB and GIGYF1 with indication of top predicted interfaces. Numbers in blue indicate the motif interface pLDDT for the respective interface. Roman numbering refers to structural models in **(B)** and **(C)**. **(B)** Structural model of interface ii shown in **(A)** (left) and in comparison a solved structure (PDB:4WZJ) of the Sm ring complex (right) bound to RNA (orange). The LSM domain of SNRPB is shown in cyan. The position of the predicted motif (left) or neighboring LSM domain of SNRPD3 (right) are indicated in gold. Black circles indicate the predicted interface in the model and corresponding interface in the complex on the LSM domain of SNRPB. **(C)** Structural model of interface i shown in **(A)** with tested domain mutations labeled and colored green. The motif sequence is indicated at the bottom. **(D, E)** BRET titration curves are shown for wildtype interactions, deletion constructs of SNRPB, and single point mutants in GIGYF1 for two biological replicates, each with three technical replicates. Protein acceptor over protein donor expression levels are plotted on the *x*-axis determined from fluorescence and luminescence measurements, respectively. **(F)** Cropped immunoblot of input (5%) and HA antibody immunoprecipitation (IP) performed in parental HEK cells (empty, untagged negative control), Snrpb(full-length, 1-231)-2xHA-mNeonGreen, Snrpb(1-190)-2xHA-mNeonGreen expressed from a single locus in Flp-In™ T-REx™ 293 Cell Lines. The HA antibody was used for detecting the immunoprecipitated Snrpb-proteins, endogenous GIGYF1 was detected with GIGYF1 antibody, GAPDH serves as a loading and negative-IP control. The experiment was performed twice with equivalent outcome, one representative experiment is shown. **(G)** Solved structure (PDB:7RUQ) of the GYF domain of GIGYF1 bound to a proline-rich motif in TNRC6C. The sequence of the motif in TNRC6C is indicated. Source data are available online for this figure.

interaction appeared less pronounced upon truncation of the C-terminal proline-containing region of SNRPB (Fig. 6F). This further suggests that both proteins interact with each other in cells and that this interaction is stabilized by the predicted interface.

During the course of these studies, a structure was published (PDB:7RUQ, Sobti et al, 2023) showing binding of the GYF domain of GIGYF1 to a motif of sequence PPPGL of the protein TNRC6C confirming the binding mode predicted by AF where a hydrophobic residue (M or L) inserts into a hydrophobic pocket and where the proline residues contact the surrounding domain surface (Fig. 6C,G). Interestingly, this hydrophobic pocket does not exist in the previously solved structure of the GYF domain of CDBP2 binding to a proline-rich peptide that is flanked by positively charged residues establishing important contacts with the domain (PDB:1L2Z, (Freund et al, 2002)). This structure formed the basis for the definition of the LIG_GYF motif class in the ELM DB. The recently resolved structure of the GYF domain of GIGYF1 together with our structural models and experimental validations argue for an extension of the existing motif definition or definition of a new motif subclass.

Discussion

AF has revolutionized the field of structural bioinformatics and has sparked much excitement about its potential to predict structures of

interacting proteins and bringing us closer to a structurally resolved protein interactome. However, from existing studies it largely remained unclear whether AF's performance depends on the type of interfaces and the length of submitted protein chains for interface prediction, which metrics perform best in identifying likely correct structural models of interfaces, how specific AF predictions are, and to which extent highly confident structural models can be experimentally corroborated. In this study, we showed that AF performs similarly well for interfaces between folded domains and interfaces formed between a folded domain and a short linear motif. Using minimal interacting regions for interface prediction we reached sensitivities of up to 80% similar to previously published work (Tsaban et al, 2022; Johansson-Åkhe et al, 2021). We thoroughly investigated AF's FPR using random domain-motif pairs and found it to be around 20%. However, asking AF to discriminate binders from non-binders when motif sequences carried one disruptive mutation, we found that prediction accuracies were close to random. This points to an important limitation in AF's ability to predict binding specificities and is in line with previous reports on AF's inability to predict the effect of mutations (Buel and Walters, 2022). Comparison of different metrics to discriminate good from bad structural models using either minimal interacting fragments or extensions revealed the average interface pLDDT for DDI models and the motif interface pLDDT for DMI models to be the most robust and best performing metrics. However, when manually inspecting AF predictions we found it useful to also consider AF's model confidence, suggesting that in the future a combination of different metrics might be even more powerful to discriminate good from bad structural models. The alignment depth has been previously reported to somewhat influence model accuracy (Bryant et al, 2022). While this feature was not investigated here, it might serve as a pre-filter to identify PPIs of high conservation for which structural modeling will likely be more successful. Interestingly, the number of residues or atoms predicted to be in contact with each other was poorly predictive, in contrast to a previous report (Bryant et al, 2022), confirming our observations that the tested AF versions in this study will always put both chains in contact with each other to create atomic contacts, and from visual inspection alone it is very challenging to tell good from bad structural models apart. Of note, observed differences in AF performance across studies likely originate both from using different benchmark datasets and different AF versions. Our study is unique in that it assesses multiple metrics on two different classes of interfaces, DMIs and DDIs, using two different AF versions. More work is needed to develop benchmark datasets of coiled-coil and disorder-disorder interfaces to also evaluate AF's performance for these modes of binding. Of note, our benchmark datasets almost exclusively consisted of structures that AF has seen in the training process. Interestingly, benchmark studies done with unseen structures reported similar sensitivities (preprint:Bret et al, 2023) indicating that AF is not strongly biased towards structures it has seen before.

We extensively explored the influence of protein fragment length on AF's performance and found that slight extensions of minimal motif sequences can improve prediction accuracies. Inspection of individual cases revealed novel information on important motif sequence context that was so far missing in corresponding motif entries at the ELM DB. However, longer disordered fragments or fragments containing ordered and large disordered regions generally decrease AF prediction accuracies as also reported in a recent preprint (preprint:Bret et al, 2023). Furthermore, optimal cutoffs for various metrics such as the model confidence decreased when using longer protein fragments, making them less robust for interface prediction with AF. When evaluating performance differences for longer and shorter protein fragments we identified three DMI pairs involving the motif classes DEG_APCC_KENBOX_2, LIG_Pex14_3, and LIG_GYF, for which, during fragment extension, a second known motif occurrence was added to the fragment. This second motif was selected by AF during interface prediction, displacing the original motif and leading to a high RMSD score. We removed these instances from the dataset when evaluating AF's performance on fragment extension but they point to biologically correct variability in AF prediction outcomes due to existing multivalency of many DMIs in protein interactions. Other work suggested that AF is able to select the stronger binder among two motif occurrences (Chang and Perez, 2023), which might at least in some cases guide AF motif selections. However, in other cases this motif preference might also hinder discovery of multivalency in PPIs. For example, the use of smaller protein fragments for the protein pair SNRPB and GIGYF1 enabled the discovery of a proline-rich repeat motif in SNRPB.

In comparison to predictions made using full length proteins (Burke et al, 2023) we found that protein fragmentation increased the probability of obtaining a high confidence interface prediction, especially for cases involving proteins with long disordered regions such as GIGYF1. For smaller and more globular proteins like the PEX proteins studied above, full length predictions can identify the right binding sites but these can be further substantiated by running additional predictions with smaller fragments. The fragmentation approach increases the number of prediction runs per protein pair from one to a couple hundred, depending on the length and modularity of both proteins. The vast majority of these fragment pairs should not interact. With a FPR of 20%, this means that more actual non-interacting than truly interacting fragment pairs will result in a high confidence prediction. A big challenge is thus to identify likely correct interface predictions among the many false ones. This is also illustrated by the prediction results that we obtained for the seven protein pairs that we followed up experimentally. Clearly, AF's general limited specificity contributes to these false predictions. We observed that additional sources of error can arise from exposed intramolecular binding sites resulting from fragmentation, incorrectly designed boundaries of folded regions, and docking of protein fragments into enzymatic pockets of metabolic enzymes or sites for metal ion, DNA, or RNA binding. It seems that AF is overall well suited to find binding pockets on folded domains. However, our work also clearly demonstrates that AF is able to correctly dock the matching partner structure into these pockets without the need for a pre-existence of both partner structures in the bound conformation contrary to other state-of-the-art docking algorithms. AF's high sensitivity with respect to intramolecular binding sites and wrongly fragmented folded regions will make it particularly hard to fully automate the fragment design process. Despite these challenges we found that recurrent interface predictions from overlapping fragments can help gain confidence in predictions, as also highlighted in a recent study (Bronkhorst et al, 2023), since we rarely observed this recurrence for likely wrong predictions.

Given the reported uncertainties in AF predictions, even for high confidence cutoffs, experimental validation is essential. The BRET assay used here has been shown in previous studies to be sensitive enough to quantify weakening of binding introduced by point mutations and to detect motif-mediated PPIs (Ebersberger et al, 2023; Trepte et al, 2018; Mo et al, 2022). Using the BRET assay, we were able to detect 11 out of 28 PPIs from the HuRI dataset. This retest rate is actually higher compared to retest rates of gold standard PPI datasets used in the past to benchmark various binary PPI assays including this BRET assay, attesting the overall detectability of PPIs from HuRI (Braun et al, 2009; Trepte et al, 2018; Choi et al, 2019). The NL and mCit fusions used in the BRET assay allowed us to monitor the expression levels of wildtype and mutant constructs, which is important to rule out loss of binding because of a destabilization of the protein. However, we cannot exclude the possibility that some expressed mutants might still be partially unfolded or mislocalized and thus, some loss of binding detected in our study could be unspecific and not the result of a specific perturbation of the predicted interface. Furthermore, preservation of binding observed for some other mutants at the predicted interface might result from the mutations not being disruptive enough and thus, do not necessarily disprove the predicted interface.

Despite these limitations, we were able to assess the validity of seven interface predictions using experimentation. We discovered a likely novel DMI type that mediates binding between PEX3 and PEX16, and proposed a model for how PEX3, PEX16, and PEX19 form a trimeric complex at the peroxisomal membrane. We also validated a variation of the LIG_GYF motif class in SNRPB that mediates binding to GIGYF1 thereby potentially connecting mRNA splicing with posttranscriptional control mechanisms. These results confirm in principle that AF is able to predict novel interface types and that it can be used to extend existing interface type definitions. However, our experimental results also highlight clear limitations of AF predictions. Our data suggests that FBXO28 and STX1B as well as STX1B and VAMP2 interact via coiled-coil interfaces but likely at higher stoichiometries and different conformations than predicted. We confirmed the binding pocket in ESRRG but not the predicted interfaces in PSMC5 and we could not substantiate interface predictions for TRIM37 and PNKP. Highly confident interface predictions were obtained for seven additional PPIs that await experimental validation. In summary, we provided experimental evidence and structural information for PPIs whose disruption is likely associated with neurodevelopmental disorders. This information can be explored in future studies aimed at delineating potential molecular mechanisms causing disease. Our study furthermore laid out clear limitations, perspectives, and future needs in AI-based structure prediction to bring us closer to a fully structurally annotated human protein interactome.

Methods

Selection of structures for DMI benchmark dataset

To gather a list of ELM classes with structural evidence and annotate their minimal interacting fragments, we downloaded a dataset of solved structures of all ELM classes from ELM DB on 08.10.2021 (ELM class version 1.4) for instances that are annotated as true positives (Kumar et al, 2022). The structures were subject to a series of manual inspections to check their validity for further analysis. First, since AlphaFold can only model the 20 standard amino acids, we excluded any structures with post-translational modifications in the motif. Second, structures that do not resolve all of the residues in a motif as curated by ELM DB were excluded. Third, we restrict our studies to only binary interactions, so DMIs that require more than two proteins to form the binding interface were excluded. Likewise, DMIs with only intramolecular interaction evidence were excluded. We manually annotated the boundaries of the domains by visual inspection of the structures. After this filtering, we identified 136 structures from distinct ELM classes that formed our DMI benchmark dataset (Dataset EV2).

Sequence identity of the domains in the DMI benchmark dataset

We took all the binding domains in the DMI benchmark dataset and computed their pairwise sequence identity from a global alignment without gap penalties. Matching residues were given a score of 1, otherwise 0. The sum of these scores was divided by the length of the longer sequence to compute the sequence identity.

Selection of structures for the DDI benchmark dataset

We randomly selected 80 pairs of Pfam domain types that were described in the 3did resource (Mosca et al, 2014) to be in contact with each other in solved structures in the Protein Data Bank (PDB). We manually inspected all PDB entries listed to contain contacts between instances of a given Pfam domain pair until we found one that we considered a genuine domain-domain interaction. These decisions were primarily based on the number of atomic contacts observed and the validity that two folded domains were interacting with each other. Out of the 80 selected Pfam domain pairs, we identified 48 DDI types and 48 corresponding approved DDI structural instances that we selected for the DDI benchmark dataset. The sequences of the minimal interacting domain regions were manually annotated by visual inspection of the structures and used for prediction. A more detailed description of the curation procedure and information on the pairs will be soon published elsewhere (Geist et al, in preparation).

Generation of random reference sets with minimal interacting regions

Mutating motif sequences

Key conserved residues of the motifs in the DMI benchmark dataset were identified computationally using the regular expression of the corresponding ELM class in the ELM DB and SLiMSearch (Krystkowiak and Davey, 2017). The defined positions are any positions in the regular expression that are not wildcards. To mutate the key residues to the ones with opposite physico-chemical properties, we substituted one or two key residues with the ones that are of the largest Miyata distance (Miyata et al, 1979) (Dataset EV2).

Randomizing pairings of known domain-motif interfaces

To simulate non-binding domain-motif pairs, we randomized the pairings of known domain motif interfaces. As some domain types can bind to motifs from distinct ELM classes, we manually checked

that the randomized pairings did not coincide with actual domain-motif interface types (Dataset EV2).

Randomizing pairings of known domain-domain interfaces

The pairings between known domain-domain interfaces were randomized to form the random reference set for DDIs.

Generation of positive DMI reference set with fragment extensions

Among the 136 solved structures that we selected previously, we further filtered for structures that consist of only human proteins. To test the potential effect of extension on DMIs that were predicted with different accuracies in their minimal forms, we selected 12 DMI types from the correct sidechain category, 8 DMI types from the correct backbone category and 11 DMI types from the correct pocket category as determined using the motif RMSD calculation. In total, 31 DMI types were selected for extension. Three additional DMI types were originally selected but later on discarded because they contained secondary motif occurrences complicating data analysis. The extensions were done on the canonical sequence of the proteins used to solve the structure. Motif extension 1 extended the motif sequence at both N and C termini by n residues where n is the length of the known motif. Motif extension 2 further extended the motif sequence by another n residues at both termini. Motif extension 3 and 4 each extended the motif sequence by $2n$ residues at both termini. Motif extension 5 extended the motif sequence by including neighboring domains and motif extension 6 used the full-length protein sequence. On the domain side, domain extension 1 extended the domain sequence to include the disordered regions N- and C-terminally of the binding domain until it reached neighboring domain(s) boundaries. Domain extension 2 included the sequence region of the neighboring domains and domain extension 3 used the full-length protein sequence. In cases where the known motif or binding domain is at the C terminus, we extended the motif or domain sequence on only the N terminus and vice versa. There were some cases where the last extension steps, motif extension 6 and domain extension 3, extended the protein minimally (<20 residues N or C terminal to the previous extension step). These cases were excluded from the analysis. The dataset of extended DMIs is in Dataset EV5. In total, 709 fragment pairs were submitted to AlphaFold. From these, 632 and 616 were successfully modeled by AF v2.2 and v2.3, respectively.

Generation of random DMI reference set with fragment extensions

To generate a random reference set using the extensions, we randomized the pairings of the 34 DMI types that we selected for extensions and paired their extensions for prediction. Motif extension 6 and domain extension 3 were excluded from the pairing. The dataset of DMIs with random pairings and their extensions can be found in Dataset EV6. In total, 612 predictions were generated, among which 566 and 522 predictions were successfully modeled by AF v2.2 and v2.3, respectively. Since motif extension 6 and domain extension 3 were excluded from the random reference set using the extensions, we also excluded them from the positive reference set extensions during ROC analysis. This resulted in 563 and 540 predictions from the positive reference set extensions for AF v2.2 and v2.3, respectively.

Selection of reference datasets for comparison of AF v2.2 with v2.3

All predictions for the minimal DMIs and the random DMIs involving minimal fragments were successfully modeled by both versions of AF. Some extensions from the positive reference set were not successfully modeled by AF v2.2 and v2.3 due to failure from HHblits. To compare AF v2.2 with v2.3, we used only predictions that were successfully modeled by both versions of AF. This resulted in 616 predictions from the extensions of the positive reference set.

Evaluation of AF sensitivity and specificity when using the fragmentation approach

Among the 34 DMIs selected for extension, we further selected 20 DMIs and retrieved the PPIs mediating these DMIs as the PRS and randomized their pairing to form random domain-motif protein pairs as the RRS. The 20 PPIs from the PRS and the 20 protein pairs from the RRS were subjected to the fragmentation approach, generating 8943 fragment pairs and 11,045 fragment pairs for the PRS and RRS, respectively. All fragment pairs from the PRS and all but one fragment pair from the RRS resulted in an AlphaFold model. Models were deemed highly confident, if the disordered fragment had a motif interface pLDDT of ≥70 or, in case of ordered-ordered models, the average interface pLDDT scored ≥70. To evaluate the sensitivity of the fragmentation approach, we considered all models that met the above mentioned cutoffs and which contained the motif and domain sequence. We superimposed the models onto the corresponding native structures using the minimal domain and computed the RMSD between the minimal motif residues in the native and modeled structure. A model was deemed accurate if the motif RMSD was ≤5 Å. At this cutoff the backbone of the native and modeled motif are well aligned but not necessarily their side chains (see also RMSD subsection below). We repeated the same procedure for each DMI protein pair using full length sequences as input into AF for modeling. In 18 cases AF did not return a model when using full length sequences. Here, we used the largest protein fragments instead for which AF returned a model. Information on the protein pairs, prediction results, and statistics is available in Dataset EV9.

AlphaFold versions and runs

We used local installations of AlphaFold Multimer version 2.2.0 and 2.3.0 (preprint:Evans et al, 2021) for all protein complex predictions with the following parameters:

--max_template_date=2020-05-14
--db_preset=full_dbs
--use_gpu_relax=False

For every AlphaFold run, five models were predicted with single seed per model by setting the following parameter:

--num_multimer_predictions_per_model=1

The databases queried during AlphaFold predictions were specified following the instructions from the github page of AlphaFold

(https://github.com/deepmind/alphafold#running-alphafold):

For running AlphaFold Multimer v2.2, the following databases were queried:

--bfd_database_path=bfd_metaclust_clu_complete_id30_c90_-final_seq.sorted_opt

--mgnify_database_path=alphafold_v220_databases/mgy_clusters_2018_12.fa

--obsolete_pdbs_path=alphafold_v220_databases/pdb_mmcif/obsolete.dat

--pdb_seqres_database_path=alphafold_v220_databases/pdb_seqres/pdb_seqres.txt

--template_mmcif_dir=alphafold_v220_databases/pdb_mmcif/mmcif_files

--uniprot_database_path=alphafold_v220_databases/uniprot/uniprot.fasta

--uniclust30_database_path=alphafold_v220_databases/uniclust30/uniclust30_2018_08/uniclust30_2018_08

--uniref90_database_path=alphafold_v220_databases/uniref90/uniref90.fasta

For running AlphaFold Multimer v2.3, the following databases were queried:

--bfd_database_path=alphafold_v230_databases/bfd/bfd_metaclust_clu_complete_id30_c90_final_seq.sorted_opt

--mgnify_database_path=alphafold_v230_databases/mgnify/mgy_clusters_2022_05.fa

--obsolete_pdbs_path=alphafold_v230_databases/pdb_mmcif/obsolete.dat

--pdb_seqres_database_path=alphafold_v230_databases/pdb_seqres/pdb_seqres.txt

--template_mmcif_dir=alphafold_v230_databases/pdb_mmcif/mmcif_files

--uniprot_database_path=alphafold_v230_databases/uniprot/uniprot.fasta

--uniref30_database_path=alphafold_v230_databases/uniref30/UniRef30_2021_03

--uniref90_database_path=alphafold_v230_databases/uniref90/uniref90.fasta

To test the effect of template use on prediction accuracy, the following parameter setting was used to switch off the use of templates during the prediction:

--max_template_date=1950-01-01

For the fragmentation approach, the multiple sequence alignments (MSAs) of a given protein fragment can be reused in subsequent runs where the same fragment is involved. The MSAs were first moved to the prediction output folder and the following parameter was added to enable the reuse of MSAs.

--use_precomputed_msas=True

For efficient computing, we segregated the MSA generation part by using only the CPUs and the model fitting part using the GPUs.

Calculation of metrics for structural models

Motif RMSD

We used the software PyMOL (TM) Molecular Graphics System, Version 2.5.0. Copyright (c) Schrodinger, LLC., for the superimposition of AlphaFold models with corresponding solved structures. First, we used the align command to align the domain chain in AlphaFold models with the domain chain in the solved structure. Then, we used the rms_cur command to calculate the all-atom RMSD between the motif chain in AlphaFold models and the motif chain in the solved structure. To ensure that the RMSD calculation was done based on all atom identifiers and without any outlier rejection refinement, the arguments of the rms_cur command, matchmaker and cycles, were set to 0. Prediction accuracy categories were defined based on motif RMSD cutoffs: RMSD ≤ 2 Å for correct sidechain, between 2 Å and 5 Å for correct backbone, between 5 Å and 15 Å for correct pocket and >15 Å for wrong pocket.

DockQ

The calculation of DockQ scores of AlphaFold models was done in reference to their solved structures using the code available on the github repository of DockQ (https://github.com/bjornwallner/DockQ, (Basu and Wallner, 2016). DockQ classification was done using the cutoffs provided by DockQ (DockQ: <0.23 for incorrect, between 0.23 and 0.49 for acceptable, between 0.49 and 0.80 for medium and ≥0.80 for high).

pDockQ

The calculation of pDockQ of AlphaFold models was done by adapting the code available on the github repository from the Elofsson lab (https://gitlab.com/ElofssonLab/FoldDock/-/blob/main/src/pdockq.py, (Bryant et al, 2022)). The pDockQ score is created by fitting a sigmoidal curve to the DockQ scores of a series of AlphaFold predicted models. The score takes into account the number of interface contacts as well as their pLDDT scores. Of note, the calculation of pDockQ score takes Cβs (Cα for glycine) from different chains within 8 Å from each other as interface contacts which is different from our interface definition (see the subsection below *Domain chain and motif chain interface pLDDT and average interface pLDDT).*

iPAE

The calculation of iPAE of AlphaFold models was done by adapting code available on the github repository https://github.com/fteufel/alphafold-peptide-receptors/tree/main (Teufel et al, 2023). The iPAE is the median predicted aligned error at the interface. The authors consider residues in contact if their distance is below 0.35 nm (3.5 Å). The iPAE score could not be calculated for models generated by AlphaFold Multimer version 2.3.0 due to JAX dependency of the pickle files generated by AlphaFold Multimer version 2.3.0.

Model confidence

The model confidence of AlphaFold models was extracted from the ranking_debug json file. The model confidence is a weighted combination of pTM and ipTM to account for both intra- and interchain confidence:

$$model\ confidence = 0.8 \cdot ipTM + 0.2 \cdot pTM$$

Domain chain and motif chain interface pLDDT and average interface pLDDT

Since AlphaFold conveniently stores the pLDDT confidence measure for each residue in the B-factor field of the output PDB files, the pLDDT of residues at the interface was parsed from the output PDB files of AlphaFold. Residues at the interface are defined as those that have at least one heavy atom that is less than 5 Å away from any heavy atom of the other chain (calculated using the

PyMOL API). The pLDDT of the residues at the interface from the domain chain and motif chain was averaged to compute the domain chain and motif chain interface pLDDT, respectively. The pLDDT of all the residues from both chains was averaged to compute the average interface pLDDT.

Residue-residue and atom-atom contacts

Following the interface definition above, the number of unique residue-residue and atom-atom contacts were also quantified as measurements to assess AlphaFold models.

Mean DockQ between predicted models

The top five models generated by AF, determined based on their model confidence, were considered for computing this metric. To quantify the similarity among the models, we computed DockQ scores between all possible pairs of models by taking the higher ranked model as the "template" model and lower ranked model as the "predicted" model. The mean of these DockQ scores is taken as the similarity among the models in a given prediction. This calculation was done for AF models of minimal DMIs and their randomizations for ROC analysis. The data were stored in Dataset EV2.

Quantification of motif properties

Motif hydropathy score and symmetry score

By referring to the Kyte-Doolittle hydrophobicity scale, (Kyte & Doolittle, 1982) the hydropathy scores of the amino acids in a given motif were summed and averaged to compute the average hydropathy of the motif. The average motif symmetry score was computed by taking the sum of the absolute difference of hydropathy scores between motif position n and n - motif length + 1 and division of this sum by half of the motif length:

$$Peptide\ symmetry\ score = \frac{\sum_{n=1}^{a}|(H_n - H_{x-n+1})|}{a}$$

where x is the length of the motif and a is the floor division of x by 2.

Motif probability

The motif probability reflects the degeneracy of a given motif class as quantified by its regular expression that is annotated in the ELM DB. The motif probability was retrieved from the ELM DB version 1.4.

Secondary structure elements of motifs

We extracted the secondary structure elements of motifs using the PyMOL API. In cases where the motif adopts partial secondary structure, such as loop-helix-loop or loop-strand-loop, they are treated as helical or strand, respectively.

Selection of motif classes from ELM DB without annotated structural instances and prediction with AF

By querying the ELM DB for all ELM classes, we retrieved a list of ELM classes and the number of instances with a structure solved (column #instances_in_PDB). We filtered for ELM classes with 0 instances_in_PDB and selected 205 instances out of the filtered ELM classes for AF prediction. The ELM instances were extended at both N and C termini by n residues where n is the length of the ELM instance, according to the benchmarking results. The minimal binding domains of the ELM instances were detected in the interaction partner using Pfam HMMs (Mistry *et al*, 2021). As the domain boundaries detected by Pfam HMMs could be inaccurate, we also extended the domain sequence at the N and C terminus by 20 residues to ensure that the whole folded region was covered. The predictions were performed using AF version 2.3.0. To select a subset of these motif classes, where we can do experimental testing, we also used the InParanoid resource (Persson & Sonnhammer, 2023) to map ELM instances where both proteins are from mouse to their human orthologs. To verify that they indeed do not have structural homologues in the PDB, we both used the SIFTS mapping (Dana *et al*, 2019) between the Pfam domain in ELM and the PDB and also looked at the ELM classes that were listed as homologs on the ELM website.

Evaluation of effect of fragment extensions on AF prediction accuracies

We superimposed the AF models generated with DMI extensions onto the corresponding solved DMI structures to quantify AF prediction accuracy using motif RMSD calculations. To this end, we aligned the two structures on their minimal binding domains and calculated the all-atom RMSD between the minimal motif in the extension AF model and the minimal motif in the solved structure. To determine potential differences in DMI prediction accuracy when using minimal versus extended protein fragments, we computed the log2 fold change of the all-atom motif RMSD before and after extension.

$$Fold\ change\ in\ prediction\ accuracy = log_2\left(\frac{all\ atom\ RMSD\ motif_{minimal\ DMI}}{all\ atom\ RMSD\ motif_{extended\ DMI}}\right)$$

Fragment design and fragment pairing for fragmentation approach

We first inspected the monomeric structural models from the AlphaFold database (Varadi et al, 2022; Jumper et al, 2021) of both interacting proteins to determine the boundaries of their ordered and coiled-coil regions, which were also treated as "ordered". All regions that were not annotated as ordered were annotated as disordered. In some cases, an extended loop with low pLDDT can be found within an ordered region. As they can also potentially carry a motif or mediate interactions in another way, these regions were also annotated as disordered in addition to their annotation as being part of a larger ordered region. The disordered regions of the proteins were fragmented into fragment sizes of 10, 20 and 30 residues. To allow AF to sample continuous sequences, we also generated another set of fragments of same sizes that overlap with the previous fragments by sliding the sequence by half the size of the fragment. The unfragmented disordered regions, as well as their fragments, from one protein were then paired with the ordered regions from its interacting partner and vice versa for prediction. The ordered regions from both proteins were also paired for prediction. We decided to manually define boundaries between ordered and disordered regions because testing available code developed for this purpose, like clustering using the PAE matrix,

turned out to be too inaccurate. We observed that erroneous removal of residues close to the domain borders that are still contributing to the folding of a structured domain, can heavily mislead AF predictions.

Selection of NDD proteins

A list of NDD genes was assembled using whole exome and whole genome sequencing studies of cohorts of NDD patients from Gene4Denovo (Zhao et al, 2020) and Deciphering Developmental Disorders (DDD) study (Firth et al, 2011), respectively. From Gene4Denovo, we selected genes linked to autism-spectrum disorders (ASD), intellectual disability (ID), epilepsy (EE), undiagnosed developmental disorders (UDD) and NDDs in general. Genes with non-coding mutations as well as genes with a false discovery rate (FDR) >= 0.05 were excluded. Similarly, in the DDD study, genes associated with developmental disorders with a neurological component, as well as genes found to be mutated in at least three children with NDDs (labeled as confirmed genes) were retained. The final list included 984 NDD-risk genes. We filtered the HuRI network (Luck et al, 2020) for interactions mediated exclusively by proteins from this NDD gene list resulting in 67 PPIs excluding self-interactions. Since our fragmentation approach generates many fragments, we did not consider PPIs involving proteins that are more than 1500 amino acids in length, resulting in a final list of 62 PPIs that were subjected to AF modeling.

Manual inspection of interface predictions for NDD-NDD PPIs and selection for experimental validation

Paired fragments from NDD-NDD PPIs were predicted using AF version 2.2 and the prediction results are stored in Dataset EV10. Based on our benchmarking results, we started by manually inspecting all NDD-NDD PPIs that obtained at least one structural model with either a motif chain interface pLDDT of ≥70 for the disordered fragment or with an average interface pLDDT ≥ 70 for structural models with predicted ordered-ordered interfaces (DDIs). However, during the course of these manual inspections, we found that using in addition a model confidence of ≥0.7 for ordered-ordered fragment pairs helped discriminating good from bad structural models. We inspected the ranked_0 models for all fragment pairs that met the above cutoffs but also inspected models scoring somewhat below these cutoffs. For every NDD-NDD PPI we used Interactome3D (Mosca et al, 2013) and PDB database searches (https://www.rcsb.org/ (Berman et al, 2000)) to identify whether a structure already existed for this PPI. In our evaluation of the structural models we also considered if a certain interface was recurrently predicted for different overlapping fragments because this usually hints at increased confidences for the correctness of the interface prediction. We furthermore explored the number and kind of residue-residue contacts predicted by AF by visual inspection of the structural models using PyMol. We searched for functional annotations and existing structures for the monomers using the PDB, ProViz (Jehl et al, 2016), SMART (Letunic et al, 2021), and the scientific literature to identify enzymatic pockets or binding interfaces for DNA, RNA, or metal ions. Observations and justifications for the final evaluation of the predictions for every NDD-NDD PPI are provided in Appendix Supplementary Text S1.

Based on clone availability, we selected 49 of the 62 PPIs for experimental validation of the predicted interfaces using the BRET assay. For 30 of the 49 selected PPIs for experimental testing we obtained sequence-confirmed clones with luciferase and mCitrine fusions. For 28 of these PPIs both partners were expressed in our experimental system as determined by total luminescence and fluorescence measurements (Fig. 3D,F).

Softwares used

We used the software PyMOL (TM) Molecular Graphics System, Version 2.5.0. Copyright (c) Schrodinger, LLC., for the visualization and superimposition of AlphaFold models.

All codes were written in Python3 and analyses were done using Jupyter notebooks. We used the Python libraries, Biopython (Cock et al, 2009) for sequence similarity computation, pandas (McKinney, 2010) for data analysis, and Matplotlib (Hunter, 2007) and seaborn (Waskom, 2021) for data visualization. ROC and PR statistics were calculated using the Python package sci-kit learn (Pedregosa et al, 2012).

Cell line culture and maintenance

HEK293 cells were purchased from DSMZ (catalog number ACC305). These cells were grown and maintained in DMEM (Thermo Fisher), supplemented with 10% FBS (PAN-Biotech), 2 mM glutamine (Thermo Fisher) and 1% penicillin–streptomycin (Thermo Fisher). Cells were incubated at 37 °C with 5% CO_2. Subcultivation was performed with 1 ml of 0.05% trypsin every 2–3 days for up to 40 passages. For each passage $1–2 \times 10^6$ cells were seeded in T25 flasks (Sarstedt). Then, new cells were thawed from stocks containing 2×10^6 cells in 1 ml of growth medium, supplemented with 10% DMSO (Sigma). Every 3 months cells were checked for mycoplasma contamination using a PCR test (Dataset EV11). The cell line was purchased from DSMZ four years ago, expanded, aliquoted, and frozen. A new aliquot is thawed after every 40 passages. No further authentication of the cell line has been done.

Plasmid construction

Standard controls

The donor and acceptor vectors pcDNA3.1-cmyc-NL-GW (Addgene plasmid ID #113446), pcDNA3.1-GW-NL-cmyc (Addgene plasmid ID #113447), pcDNA3.1 GW-His3C-mCit, pcDNA3.1 mCit-His3C-GW as well as controls pcDNA3.1-NL-cmyc (Addgene plasmid ID #113442), pcDNA3.1-PA-mCit (Addgene plasmid ID #113443) were kindly provided by the Wanker Group (Max-Delbrück-Centrum für Molekulare Medizin, Germany) (Dataset EV12). By default we cloned all ORFs of interest into N-terminal NL and mCit fusion destination vectors and occasionally also transferred ORFs into C-terminal fusion vectors if N-terminal fusions did not result in sufficient BRET signals but the interaction was of high interest to this study and predicted interfaces were closer to the C-terminus. Trepte et al have shown that testing protein pairs in different configurations increases detection rates while maintaining low false detection rates and that BRET signals are higher if fusions are close to the actual interaction interface (Trepte et al, 2018; preprint:Trepte et al, 2021; preprint:Trepte et al, 2023).

GATEWAY cloning procedure

Full-length wild-type human open reading frames (ORFs) being cloned in GATEWAY entry vectors from the ORFeome collaboration are stored as bacterial glycerol stocks. (ORFeome Collaboration, 2016)

1. The ORFs were inoculated in 96-well plates (Corning), with each well containing 200 uL of LB medium and 100 µg/ml ampicillin. The plate was incubated at 37 °C and left to shake overnight at 190 rpm.
2. In a 96-well PCR plate (Brand) 10 ng of each selected ORF was used per 50 µl PCR reaction (denaturation at 98 °C for 10 s, annealing at 55 °C for 30 s and extension at 72 °C for 3 min, 30 cycles of amplification) using phusion high-fidelity polymerase (NEB) and primers annealing to the backbone of the plasmid (forward: 5′TTGTAAAACGACGGCCAGTC and reverse: 5′ GCCAGGAAACAGCTATGACC).
3. The PCR products (6 µl per well) were confirmed through 96-well E-gel with SYBR (Thermo Fisher, Catalog no G720801) using 25 µl of loading buffer (Thermo Fisher) and 20 µl of E-Gel 96 High range DNA marker (Thermo Fisher).
4. In a 96-well PCR plate 1 µl of each amplified PCR product together with 200 ng of above-mentioned destination vectors were directly used per 10 µl LR reaction using 4x LR clonase (Invitrogen), thereby generating expression vectors.
5. The full 10 µl of LR reaction was transformed into chemically competent DH5a cells (30 µl) in a 96-well PCR plate, then recovered in 80 µl of pre-warmed SOC medium at 37 ˚C for 1 h without shaking.
6. 70 µl of transformed bacteria was plated on 48-well square agar plates and incubated at 37 °C overnight.
7. Afterwards, colonies were selected and inoculated into a 96 deep-well plate containing 2 ml of LB medium and 100 µg/ml ampicillin. The plate was then incubated at 37 ˚C with continuous shaking at 700 rpm in the incumixer for 24 h.
8. The amplified vectors were extracted from the inoculated culture using Plasmid Plus 96-well Miniprep kit (Qiagen). The concentration of each vector was measured with a Nanophotometer and diluted to 100 ng/µl. Next, 600 ng of insert was used for full-length sequencing using the backbone primers (tag-specific NanoLuc forward: 5′GAACGGCAACAAAATTATC-GAC, mCitrine forward: 5′AGCAGAATACGCCCATCG and reverse: 5′GGCAACTAGAAGGCACAGTC) and ORF-specific primers (Dataset EV11) to fully cover the ORFs where it was needed (Dataset EV12). All sequence-confirmed ORF sequences used in this study are available in Dataset EV13.

Site-directed mutagenesis

The primers were manually designed using the following criteria:

1. For point mutation the primers should overlap the site of mutation. The overlap should be 15–20 nucleotides (nt).
2. For the deletion the primers should be designed to exclude the deletion site, but still overlap and the overlap should be as mentioned in step 1.
3. Primer length should be in the range of 32–36 nt.
4. GC content should be between 40–60%.
5. Difference in melting temperature of primers should not exceed 5 °C.
6. The primer ideally should start and end with guanine or cytosine.
7. The designed oligos were grouped by annealing temperature for the next step.
8. In 96-well PCR plate 10 ng of DNA template together with oligos were used per 50 µL of PCR reaction (denaturation at at 98 °C for 2 min, annealing for 15 s and extension at 72 °C for 5 min, 25 cycles of amplification) using phusion high-fidelity polymerase (NEB).
9. 1 µL of DpnI (NEB) was added to the plate with PCR products and incubated at 37 °C for 1 h. The reaction was stopped at 65 °C for 20 min.
10. The PCR products (6 µl per well) were confirmed through 96-well E-gel with SYBR (Thermo Fisher, Catalog no G720801) using 25 µl of loading buffer (Thermo Fisher) and 20 µl of E-Gel 96 High range DNA marker (Thermo Fisher).
11. 3 µL of digested PCR product was transformed into chemically competent DH5a cells (30 µL) in a 96-well PCR plate, then recovered in 80 µL of pre-warmed SOC medium at 37 °C for 1 h without shaking.
12. 70 µL of transformed bacteria was plated on 48-well square agar plates and incubated at 37 °C overnight.
13. Afterwards, colonies were selected and inoculated into a 96 deep-well plate containing 2 ml of LB medium and 100 µg/ml ampicillin. The plate was then incubated at 37˚C with continuous shaking at 700 rpm in the incumixer for 24 h.
14. The amplified vectors were extracted from the inoculated culture with Plasmid Plus 96-well Miniprep kit (Qiagen). The concentration was measured with a Nanophotometer and diluted to 100 ng/µl. Next, 600 ng of insert was used for full-length sequencing using primers covering the mutation and ORF-specific primers (Dataset EV11) to fully cover the ORF length (Dataset EV12).

BRET assay

Transfection

HEK293 cells were grown and maintained in high-glucose (4.5 g/l) DMEM (Thermo Fisher) for BRET assays. Media was supplemented with 10% fetal bovine serum (PAN-Biotech) and 1% Penicillin/Streptomycin. Cells were grown at 37 °C, 5% CO_2, and 85% RH. Cells were subcultured every 2–3 days and transfected with lipofectamine 2000 transfection reagent (Invitrogen) in Opti-MEM medium (Thermo Fisher) using the reverse transfection method according to the manufacturer's instructions. For transfections, cells were seeded at a density of 4.0×10^4 cells per well in a white 96-well microtiter plate (Greiner) in phenol-red-free, high-glucose DMEM media (Thermo Fisher) supplemented with 5% fetal bovine serum (Thermo Fisher). Transfections were performed with a total DNA amount of 200 ng per well. If the expression plasmid concentration amount was below 200 ng/well, pcDNA3.1 (+) was used as a carrier DNA to reach the total amount of DNA of 200 ng. All protein pairs were tested in both N-terminal fusion orientations (NL-A with mCit-B and NL-B with mCit-A). The following proteins were also tested as C-terminal fusions: CSNK2B-NL, ESRRG-NL, CUL3-NL, PEX3-NL, PEX19-NL, PSMC5-NL, PEX3-mCit, PEX19-mCit, PEX16-mCit, RORB-mCit, ESRRG-mCit, PAX6-mCit, CSNK2B-mCit, PSMC5-mCit, KCTD7-mCit (Dataset EV12).

Measurement

The plate was incubated 2 days at 37 °C, 5% CO_2, and 85% RH before measurements. All measurements were done with the Infinite M200 Pro microplate reader (Tecan). First, 100 µl of the medium was aspirated from each well. The mCitrine fluorescence (FL) was measured in intact cells (excitation/emission 513 nm/548 nm) using a gain of 100. On rare occasions, the plate reader recorded an overflow with these settings (i.e. for GIGYF1 constructs). In these cases, we repeated the measurement with optimal gain settings and used a fluorescein control to normalize fluorescence signals measured with different gain settings. For this purpose, Fluorescein was obtained from Sigma-Aldrich (Catalog No 46955-250MG-F) and used without further purification. A stock solution of Fluorescein (1 mg/ml in Ethanol) was prepared by dissolving 1.3 mg Fluorescein in 1.3 ml absolute ethanol. 100 µl of a 20 µg/ml solution of Fluorescein were added to an empty well immediately before starting the fluorescence measurements. The 20 µg/ml solution of Fluorescein was obtained by preparing a 1:50 dilution in water of the stock solution. After measuring the fluorescence, coelenterazine-h (PJK Biotech GmbH) was added to a final concentration of 5 µM. The cells were briefly shaken for 15 s and incubated for 15 min inside the plate reader at 37 °C. After incubation, total luminescence was measured first followed by short-wavelength (WL) and long-wavelength luminescence (LU) measurements using the BLUE1 (370–480 nm) and the GREEN1 (520–570 nm) filters at 1000 ms integration time. Corrected BRET ratios were calculated as described in (Trepte et al, 2018). Briefly, for every transfected protein pair NL-A and mCit-B, the following two control pairs were measured: NL-Stop with mCit-B and NL-A with mCit-Stop. The maximal BRET from both control pairs was subtracted from the actual test pair to correct for donor bleedthrough, unspecific binding to the tags, and background signal.

Determination of binding events in BRET assay

To determine whether a protein pair interacted in the BRET assay or not, we used donor:acceptor DNA transfection ratios of 2:50 ng in all cases except for PEX3-PEX16 where we used 8:25 and PEX3:PEX19 where we used 8:50 ng DNA ratios due to low expression levels of PEX3 and a degradation effect of higher PEX16 protein levels on PEX3 expression levels. We requested that cBRETs determined at these transfection ratios were ≥0.05, fluorescence measurements representing mCitrine fusion expression levels to be ≥500 units, and total luminescence measurements representing NL fusion expression levels to be ≥50,000.

Saturation assay

For donor saturation experiments various donor DNA amounts (1, 2, 4 and 8 ng) encoding NL-fused proteins were co-transfected with increasing amounts of acceptor DNA (12.5, 25, 50, 100, 200 ng) encoding mCitrine-fused proteins. Fluorescence, total luminescence, and BRET measurements were done as described before. BRET measurements were corrected for bleedthrough using NL-Stop transfections. Fluorescence and total luminescence measurements were corrected for background signal using transfections with pcDNA3.1(+) and subsequently used to estimate amounts of expressed proteins and to plot acceptor/donor ratios on the *x*-axis of titration plots.

Fitting of titration curves

Titration curves were fitted using the leastsq function from the scipy.optimize python package (Virtanen et al, 2020) using the model BRET = ((A/D) * BRETmax)/(BRET50 + (A/D)) described in (Drinovec et al, 2012), which assumes a 1:1 binding mode, to obtain estimates for the BRETmax and BRET50. Standard errors of the BRET50 estimates were obtained from the variance-covariance matrix, calculated by multiplying the fractional covariance matrix (output by leastsq function) by the residual variance. Measuring BRET signals in intact cells for increasing acceptor/donor protein expression ratios results in an eventual saturation of the signal. Fitting this curve allows extraction of the maximal BRET that can be reached and the BRET50, which is the acceptor/donor ratio at which half of the maximal BRET is obtained. The BRET50 is indicative of binding affinity, in analogy to the IC50, however, its accurate estimation requires saturation of the BRET to be observed in the experimental system, which cannot always be achieved because of limited amounts of DNA that cells can be transfected with. Alternatively, if mutations are unlikely to change the overall structure of the fusion constructs and do not alter expression levels compared to wildtype, single point BRET measurements at acceptor/donor ratios prior to BRET saturation are also indicative of changes in binding strength. The BRET titration curves that we obtained for the PNKP-TRIM37 interaction clearly deviated from the assumed 1:1 binding mode because at higher acceptor:donor ratios we observed a sudden increase in BRET again contrary to an expected saturation. The model could thus not be fitted to the titration data.

Antibodies

Purified anti-HA.11 Epitope Tag, Clone: [16B12], Mouse, Monoclonal (Biolegend, BLD-901502), 1:2000.

Purified anti-GIGYF1, Rabbit, Polyclonal (BETHYL laboratories, Cat. #A304-132A-1), 1:1000.

GAPDH Loading Control Monoclonal Antibody (GA1R), HRP-coupled (Thermo Fisher Cat. MA515738HRP), 1:3000.

Co-immunoprecipitation and western blot

Snrpb (full-length) and C-terminal truncation mutant (amino acids 1-190) was cloned from mouse cDNA and ligated into pFRT-TO destination plasmid using AscI and PacI restriction sites. The constructs additionally contain C-terminal 2xHA and mNeonGreen tags. Flp-In™ T-REx™ 293 Cell Lines (Thermo Fisher, catalog number: R78007) expressing Snrpb endogenously from a single locus were generated according to the manufacturer's instructions. In brief, pFRT-TO and pOG44 plasmids were co-transfected and hygromycin-resistant colonies were grown, picked and expanded. The Snrpb transgene expression was validated by western blot, RT-qPCR, and immunofluorescence, which showed that ectopic Snrpb-HA was expressed at levels highly similar to the endogenous Snrpb protein.

For the co-immunoprecipitation experiments, 8 × 106 cells were seeded in a 10 cm dish. The following day, expression of Snrpb-HA was induced by adding 0.1 µg/mL Doxycycline (D9891, Sigma Aldrich) to the culture medium. Parental cells not expressing any HA-tagged transgene were used as a negative control of immunoprecipitation. The next morning the cells were harvested by scraping in culture media, followed by centrifugation and a

single wash in ice-cold PBS. The whole cell extract was prepared by 15 min incubation on ice with 0.3 mL of lysis buffer (200 mM NaCl, 50 mM HEPES, pH 7.6, 0.1% IGEPAL, 10 mM $MgCl_2$, 10% Glycerol, Protease Inhibitor Cocktail (P8340, Sigma Aldrich), Phosphatase Inhibitor (P5726, Sigma Aldrich) followed by 2 cycles of sonication in a Bioruptor Plus (30 s on, 30 s off) and centrifugation for 20 min at 16,000 × *g*. The extract was quantified by a Bradford assay and 1 mg was used for immunoprecipitation, for which the NaCl concentration was adjusted to 100 mM final concentration by diluting with an equal volume of Lysis Buffer containing 0 mM NaCl. 0.05 mg was set aside as input control (5%). 0.02 mL of Thermo Scientific™ Pierce™ Anti-HA Magnetic Beads (Thermo Fisher Cat. 13464229) were incubated with 1 mg protein extract for 1 h at 4 °C on a rotating wheel. The beads were washed three times before eluting the immunoprecipitated proteins with 0.02 mL of 1 x NuPAGE™ LDS Sample Buffer by incubating at 42 °C for 10 min while shaking at 800 rpm. Another 0.01 mL were used for elution, were then combined making a total of 30 µL, which were transferred to a fresh tube and to which 3 µL of 1 M DTT were added. Input and immunoprecipitated eluates were then separated on a 10% Tris-Glycine SDS PAGE using 1xMOPS buffer, immunoblotted on 0.45 µm PVDF membranes (Tris-Glycin Transfer Buffer, 10% Methanol, 300 mA, 1 hour), blocked with 5% milk in TBS-0.2% Tween for 30 min at RT. Primary antibodies were incubated overnight at 4 °C on a rocker followed by washes and incubation with secondary HRP-labeled antibodies (1 h at RT in 5% milk, TBS-0.2% Tween). Blots were developed using Pierce™ ECL Western Blotting Substrate (Thermo Fisher Cat. 32209) or SuperSignal West Femto Maximum Sensitivity Substrate Kit (Thermo Fisher Cat. 34095) and imaged on a ChemiDoc MP V3 (Bio-Rad). The cell line was authenticated via X-Gal staining, qPCR and Sanger Sequencing.

Data availability

The datasets and computer code produced in this study are available in the following databases:

- Interaction data: submitted to the IMEx (http://www.imexconsortium.org) consortium through IntAct (Del Toro et al, 2022) and assigned the identifier IM-29904.
- Computer scripts for data processing and analysis: available at GitHub under https://github.com/KatjaLuckLab/AlphaFold_manuscript.

Expanded view data, supplementary information, appendices are available for this paper at https://doi.org/10.1038/s44320-023-00005-6.

Peer review information

A peer review file is available at https://doi.org/10.1038/s44320-023-00005-6

References

Ajuh P, Chusainow J, Ryder U, Lamond AI (2002) A novel function for human factor C1 (HCF-1), a host protein required for herpes simplex virus infection, in pre-mRNA splicing. EMBO J 21:6590–6602

Akdel M, Pires DEV, Pardo EP, Jänes J, Zalevsky AO, Mészáros B, Bryant P, Good LL, Laskowski RA, Pozzati G et al (2022) A structural biology community assessment of AlphaFold2 applications. Nat Struct Mol Biol 29:1056–1067

Basu S, Wallner B (2016) DockQ: a quality measure for protein-protein docking models. PLoS ONE 11:e0161879

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28:235–242

Braun P, Tasan M, Dreze M, Barrios-Rodiles M, Lemmens I, Yu H, Sahalie JM, Murray RR, Roncari L, de Smet AS, Venkatesan K, Rual JF, Vandenhaute J, Cusick ME, Pawson T, Hill DE, Tavernier J, Wrana JL, Roth FP, Vidal M (2009) An experimentally derived confidence score for binary protein-protein interactions. Nature Methods 6:91–97

Bret H, Andreani J, Guerois R (2023) From interaction networks to interfaces: Scanning intrinsically disordered regions using AlphaFold2. Preprint at BioRxiv https://doi.org/10.1101/2023.05.25.542287

Bronkhorst AW, Lee CY, Möckel MM, Ruegenberg S, de Jesus Domingues AM, Sadouki S, Piccinno R, Sumiyoshi T, Siomi MC, Stelzl L, Luck K, Ketting RF (2023) An extended Tudor domain within Vreteno interconnects Gtsf1L and Ago3 for piRNA biogenesis in Bombyx mori. EMBO J 42(24):e114072 https://doi.org/10.15252/embj.2023114072

Bryant P, Pozzati G, Elofsson A (2022) Improved prediction of protein-protein interactions using AlphaFold2. Nat Commun 13:1265

Buel GR, Walters KJ (2022) Can AlphaFold2 predict the impact of missense mutations on structure? Nat Struct Mol Biol 29:1–2

Bugge K, Brakti I, Fernandes CB, Dreier JE, Lundsgaard JE, Olsen JG, Skriver K, Kragelund BB (2020) Interactions by disorder - a matter of context. Front Mol Biosci 7:110

Burke DF, Bryant P, Barrio-Hernandez I, Memon D, Pozzati G, Shenoy A, Zhu W, Dunham AS, Albanese P, Keller A et al (2023) Towards a structurally resolved human protein interaction network. Nat Struct Mol Biol 30:216–225

Chang L, Perez A (2023) Ranking peptide binders by affinity with AlphaFold. Angew Chem Int Ed 62:e202213362

Choi SG, Olivet J, Cassonnet P, Vidalain PO, Luck K, Lambourne L, Spirohn K, Lemmens I, Dos Santos M, Demeret C, Jones L, Rangarajan S, Bian W, Coutant EP, Janin YL, van der Werf S, Trepte P, Wanker EE, De Las Rivas J, Tavernier J, Twizere JC, Hao T, Hill DE, Vidal M, Calderwood MA, Jacob Y (2019) Maximizing binary interactome mapping with a minimal number of assays. Nature Communications 10:3907

Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B et al (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25:1422–1423

Dana JM, Gutmanas A, Tyagi N, Qi G, O'Donovan C, Martin M, Velankar S (2019) SIFTS: updated structure integration with function, taxonomy and sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. Nucleic Acids Res 47:D482–D489

Davey NE, Van Roey K, Weatheritt RJ, Toedt G, Uyar B, Altenberg B, Budd A, Diella F, Dinkel H, Gibson TJ (2012) Attributes of short linear motifs. Mol Biosyst 8:268–281

Del Toro N, Shrivastava A, Ragueneau E, Meldal B, Combe C, Barrera E et al (2022) The IntAct database: efficient access to fine-grained molecular interaction data. Nucleic Acids Res 50(D1):D648–53

Drew K, Lee C, Huizar RL, Tu F, Borgeson B, McWhite CD, Ma Y, Wallingford JB, Marcotte EM (2017) Integration of over 9000 mass spectrometry experiments builds a global map of human protein complexes. Molecular Systems Biology 13:932

Drinovec L, Kubale V, Nøhr Larsen J, Vrecl M (2012) Mathematical models for quantitative assessment of bioluminescence resonance energy transfer:

application to seven transmembrane receptors oligomerization. Front Endocrinol 3:104

Durocher D, Taylor IA, Sarbassova D, Haire LF, Westcott SL, Jackson SP, Smerdon SJ, Yaffe MB (2000) The Molecular Basis of FHA Domain:Phosphopeptide Binding Specificity and Implications for Phospho-Dependent Signaling Mechanisms. Molecular Cell 6:1169–1182

Ebersberger S, Hipp C, Mulorz MM, Buchbender A, Hubrich D, Kang HS, Martínez-Lumbreras S, Kristofori P, Sutandy FXR, Llacsahuanga Allcca L, Schönfeld J, Bakisoglu C, Busch A, Hänel H, Tretow K, Welzel M, Di Liddo A, Möckel MM, Zarnack K, Ebersberger I, Legewie S, Luck K, Sattler M, König J (2023) FUBP1 is a general splicing factor facilitating 3′ splice site recognition and splicing of long introns. Molecular Cell 83:2653–2672

Ernst JA, Brunger AT (2003) High Resolution Structure Stability and Synaptotagmin Binding of a Truncated Neuronal SNARE Complex. Journal of Biological Chemistry 278:8630–8636

Evans R, O'Neill M, Pritzel A, Antropova N, Senior AW, Green T, Žídek A, Bates R, Blackwell S, Yim J et al (2021) Protein complex prediction with AlphaFold-Multimer. Preprint at BioRxiv https://doi.org/10.1101/2021.10.04.463034

Firth HV, Wright CF, DDD Study (2011) The deciphering developmental disorders (DDD) study. Dev Med Child Neurol 53:702–703

Freiman RN, Herr W (1997) Viral mimicry: common mode of association with HCF by VP16 and the cellular protein LZIP. Genes Dev 11:3122–3127

Freund C, Kühne R, Yang H, Park S, Reinherz EL, Wagner G (2002) Dynamic interaction of CD2 with the GYF and the SH3 domain of compartmentalized effector molecules. EMBO J 21:5985–5995

Fujiki Y, Matsuzono Y, Matsuzaki T, Fransen M (2006) Import of peroxisomal membrane proteins: the interplay of Pex3p- and Pex19p-mediated interactions. Biochim Biophys Acta 1763:1639–1646

Fujiki Y, Okumoto K, Honsho M, Abe Y (2022) Molecular insights into peroxisome homeostasis and peroxisome biogenesis disorders. Biochim Biophys Acta Mol Cell Res 1869:119330

Henrie A, Hemphill SE, Ruiz-Schultz N, Cushman B, DiStefano MT, Azzariti D, Harrison SM, Rehm HL, Eilbeck K (2018) ClinVar Miner: demonstrating utility of a Web-based tool for viewing and filtering ClinVar data. Hum Mutat 39:1051–1060

Hunter JD (2007) Matplotlib: a 2D graphics environment. Comput Sci Eng 9:90–95

Huttlin EL, Bruckner RJ, Navarrete-Perea J, Cannon JR, Baltier K, Gebreab F, Gygi MP, Thornock A, Zarraga G, Tam S et al (2021) Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. Cell 184:3022–3040.e28

Jehl P, Manguy J, Shields DC, Higgins DG, Davey NE (2016) ProViz-a web-based visualization tool to investigate the functional and evolutionary features of protein sequences. Nucleic Acids Res 44:W11–5

Johansson-Åkhe I, Mirabello C, Wallner B (2021) Interpeprank: assessment of docked peptide conformations by a deep graph network. Front Bioinform 1:763102

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A et al (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596:583–589

Krystkowiak I, Davey NE (2017) SLiMSearch: a framework for proteome-wide discovery and annotation of functional modules in intrinsically disordered regions. Nucleic Acids Res 45:W464–W469

Kumar M, Michael S, Alvarado-Valverde J, Mészáros B, Sámano-Sánchez H, Zeke A, Dobson L, Lazar T, Örd M, Nagpal A et al (2022) The Eukaryotic Linear Motif resource: 2022 release. Nucleic Acids Res 50:D497–D508

Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157:105–132

Letunic I, Khedkar S, Bork P (2021) SMART: recent updates, new developments and status in 2020. Nucleic Acids Res 49:D458–D460

Leung AKW, Nagai K, Li J (2011) Structure of the spliceosomal U4 snRNP core domain and its implication for snRNP biogenesis. Nature 473:536–539

Lu R, Yang P, O'Hare P, Misra V (1997) Luman, a new member of the CREB/ATF family, binds to herpes simplex virus VP16-associated host cellular factor. Mol Cell Biol 17:5117–5126

Luck K, Charbonnier S, Travé G (2012) The emerging contribution of sequence context to the specificity of protein interactions mediated by PDZ domains. FEBS Lett 586:2648–2661

Luck K, Kim D-K, Lambourne L, Spirohn K, Begg BE, Bian W, Brignall R, Cafarelli T, Campos-Laborie FJ, Charloteaux B et al (2020) A reference map of the human binary protein interactome. Nature 580:402–408

Machida YJ, Machida Y, Vashisht AA, Wohlschlegel JA, Dutta A (2009) The deubiquitinating enzyme BAP1 regulates cell growth via interaction with HCF-1. J Biol Chem 284:34179–34188

Matsuzaki T, Fujiki Y (2008) The peroxisomal membrane protein import receptor Pex3p is directly transported to peroxisomes by a novel Pex19p- and Pex16p-dependent pathway. J Cell Biol 183:1275–1286

McKinney W (2010) Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference pp 56–61. SciPy

Mishra M, Jiang H, Wei Q (2023) New insights on the differential interaction of sulfiredoxin with members of the peroxiredoxin family revealed by protein-protein docking and experimental studies. Eur J Pharmacol 954:175873

Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ et al (2021) Pfam: the protein families database in 2021. Nucleic Acids Res 49:D412–D419

Miyata T, Miyazawa S, Yasunaga T (1979) Two types of amino acid substitutions in protein evolution. J Mol Evol 12:219–236

Mo X, Niu Q, Ivanov AA, Tsang YH, Tang C, Shu C, Li Q, Qian K, Wahafu A, Doyle SP, Cicka D, Yang X, Fan D, Reyna MA, Cooper LAD, Moreno CS, Zhou W, Owonikoko TK, Lonial S, Khuri FR, Du Y, Ramalingam SS, Mills GB, Fu H (2022) Systematic discovery of mutation-directed neo-protein-protein interactions in cancer. Cell 185:1974–1985

Mosca R, Céol A, Aloy P (2013) Interactome3D: adding structural details to protein networks. Nat Methods 10:47–53

Mosca R, Céol A, Stein A, Olivella R, Aloy P (2014) 3did: a catalog of domain-based interactions of known three-dimensional structure. Nucleic Acids Res 42:D374–9

O'Reilly FJ, Graziadei A, Forbrig C, Bremenkamp R, Charles K, Lenz S, Elfmann C, Fischer L, Stülke J, Rappsilber J (2023) Protein complexes in cells by AI-assisted structural proteomics. Mol Syst Biol 19:e11544

ORFeome Collaboration (2016) The ORFeome Collaboration: a genome-scale human ORF-clone resource. Nat Methods 13:191–192

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Müller A, Nothman J, Louppe G et al (2012) Scikit-learn: Machine Learning in Python. arXiv

Persson E, Sonnhammer ELL (2023) InParanoiDB 9: ortholog groups for protein domains and full-length proteins. J Mol Biol 435:168001

Pozzati G, Zhu W, Bassot C, Lamb J, Kundrotas P, Elofsson A (2022) Limits and potential of combined folding and docking. Bioinformatics 38:954–961

Schmidt F, Treiber N, Zocher G, Bjelic S, Steinmetz MO, Kalbacher H, Stehle T, Dodt G (2010) Insights into peroxisome function from the structure of PEX3 in complex with a soluble fragment of PEX19. J Biol Chem 285:25410–25417

Sobti M, Mead BJ, Stewart AG, Igreja C, Christie M (2023) Molecular basis for GIGYF–TNRC6 complex assembly. RNA 29:724–734

Teufel F, Refsgaard JC, Kasimova MA, Deibler K, Madsen CT, Stahlhut C, Grønborg M, Winther O, Madsen D (2023) Deorphanizing peptides using structure prediction. J Chem Inf Model 63:2651–2655

Tompa P, Davey NE, Gibson TJ, Babu MM (2014) A million peptide motifs for the molecular biologist. Mol Cell 55:161–169

Trepte P, Kruse S, Kostova S, Hoffmann S, Buntru A, Tempelmeier A, Secker C, Diez L, Schulz A, Klockmeier K et al (2018) LuTHy: a double-readout bioluminescence-based two-hybrid technology for quantitative mapping of protein-protein interactions in mammalian cells. Mol Syst Biol 14:e8071

Trepte P, Secker C, Choi SG, Olivet J, Ramos ES, Cassonnet P, Golusik S, Zenkner M, Beetz S, Sperling M et al (2021) A quantitative mapping approach to identify direct interactions within complexomes. Preprint at BioRxiv https://doi.org/10.1101/2021.08.25.457734

Trepte P, Secker C, Kostova S, Maseko SB, Choi SG, Blavier J, Minia I, Ramos ES, Cassonnet P, Golusik S et al (2023) AI-guided pipeline for protein-protein interaction drug discovery identifies a SARS-CoV-2 inhibitor. Preprint at BioRxiv https://doi.org/10.1101/2023.06.14.544560

Tsaban T, Varga JK, Avraham O, Ben-Aharon Z, Khramushin A, Schueler-Furman O (2022) Harnessing protein folding neural networks for peptide-protein docking. Nat Commun 13:176

Van Roey K, Gibson TJ, Davey NE (2012) Motif switches: decision-making in cell regulation. Curr Opin Struct Biol 22:378–385

Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A et al (2022) AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res 50:D439–D444

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J et al (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 17:261–272

Waskom M (2021) seaborn: statistical data visualization. JOSS 6:3021

Weatheritt RJ, Jehl P, Dinkel H, Gibson TJ (2012) iELM-a web server to explore short linear motif-mediated interactions. Nucleic Acids Res 40:W364–W369

Zhao G, Li K, Li B, Wang Z, Fang Z, Wang X, Zhang Y, Luo T, Zhou Q, Wang L et al (2020) Gene4Denovo: an integrated database and analytic platform for de novo mutations in humans. Nucleic Acids Res 48:D913–D926

Acknowledgements

We thank all members of the Luck, Gibson, and Schueler-Furman labs as well as Julian König and Anton Khmelinskii for helpful discussions and input. We thank Izabella Krystkowiak and Norman Davey for helping us access the SLiMSearch resource with an API. We thank Fridolin Kielisch for advice on statistical analysis as well as the media lab and protein production core facilities of IMB. Support from IMB's IT department and especially help from Christian Dietrich for local installations of AlphaFold is gratefully acknowledged. The GPU cluster on which part of the AlphaFold predictions were performed was funded by the Ministry of Science and Health (MWG), Rhineland Palatinate (funding ID: TB-Nr.:3658/19). We are very thankful for support from EMBL IT Services and the HPC resources for running AlphaFold predictions for this project. This work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-IDs LU 2568/1-1 and SFB1551 Project No 464588647 awarded to KL. JS acknowledges a PhD stipend from IMB's collaborative research initiative. JKV was supported by the European Union's Horizon 2020 UBIMOTIF programme (860517). This work was supported, in whole or in part, by the Israel Science Foundation, founded by the Israel Academy of Science and Humanities (grant number 301/2021 to OS-F).

Author contributions

Chop Yan Lee: Data curation; Formal analysis; Investigation; Visualization; Methodology; Writing—original draft; Project administration; Writing—review and editing. **Dalmira Hubrich**: Data curation; Formal analysis; Investigation; Visualization; Methodology; Writing—original draft; Writing—review and editing. **Julia K Varga**: Data curation; Formal analysis; Investigation; Visualization; Writing—original draft; Writing—review and editing. **Christian Schäfer**: Data curation; Investigation; Methodology. **Mareen Welzel**: Investigation. **Eric Schumbera**: Methodology. **Milena Djokic**: Data curation. **Joelle M Strom**: Formal analysis; Investigation; Visualization. **Jonas Schönfeld**: Investigation. **Johanna L Geist**: Investigation. **Feyza Polat**: Investigation. **Toby J Gibson**: Resources; Supervision; Writing—review and editing. **Claudia Isabelle Keller Valsecchi**: Supervision; Funding acquisition; Investigation; Writing—review and editing. **Manjeet Kumar**: Resources; Formal analysis; Methodology; Writing—review and editing. **Ora Schueler-Furman**: Conceptualization; Supervision; Funding acquisition; Writing—original draft; Writing—review and editing. **Katja Luck**: Conceptualization; Data curation; Formal analysis; Supervision; Funding acquisition; Investigation; Visualization; Methodology; Writing—original draft; Project administration; Writing—review and editing.

Disclosure and competing interest statement

The authors declare no competing interests.

2.4.1 Supplementary material

Appendix

Systematic discovery of protein interaction interfaces using AlphaFold and experimental validation

Chop Yan Lee[1,†], Dalmira Hubrich[1,†], Julia K. Varga[2,†], Christian Schäfer[1], Mareen Welzel[1], Eric Schumbera[3], Milena Đokić[1], Joelle M. Strom[1], Jonas Schönfeld[1], Johanna L. Geist[1], Feyza Polat[1], Toby J. Gibson[4], Claudia Isabelle Keller Valsecchi[1], Manjeet Kumar[4], Ora Schueler-Furman[2,*], Katja Luck[1,**]

Affiliations
1 Institute of Molecular Biology (IMB) gGmbH, 55128 Mainz, Germany.
2 Department of Microbiology and Molecular Genetics,Institute for Biomedical Research Israel-Canada, Faculty of Medicine, The Hebrew University of Jerusalem, Jerusalem 9112001, Israel.
3 Institute of Molecular Biology (IMB) gGmbH, 55128 Mainz, Germany. Current address: Computational Biology and Data Mining Group Biozentrum I 55128 Mainz, Germany.
4 Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, 69117, Germany.

*Corresponding author. Tel: +972-2-675-7094, E-mail: ora.furman-schueler@mail.huji.ac.il
**Corresponding author. Tel: +49-(0)6131-3921440, E-mail: k.luck@imb-mainz.de
†These authors contributed equally to this work.

Table of content

Appendix Figure S5. Comparison of AF v2.2 and v2.3 prediction performance.

Appendix Figure S6. Performance of different metrics derived from structural models when benchmarking AF v2.3 for DMI predictions.

Appendix Figure S7. Expression and BRET50 plots for TRIM37-PNKP and ESRRG-PSMC5.

Appendix Figure S8. Structural models, expression, and BRET50 plots for STX1B-FBXO28 and STX1B-VAMP2.

Appendix Figure S9. Structural models, expression, and BRET50 plots for PEX3-PEX19 and PEX3-PEX16.

Appendix Figure S10. Expression and BRET50 plots for SNRPB-GIGYF1.

Appendix Text S1. Summary of observations from the manual inspection of AlphaFold models generated from fragmentation approach on PPIs connecting NDD proteins.

run14: PLP1-MFF

Top prediction involves an ordered region from PLP1 and a disordered fragment from MFF, with a model confidence of 0.75. Looking at the predicted model, the peptide is tilted at an angle to the bundle of helices of PLP1, not like the usual coiled-coil interaction. No trend in increasing confidence with shorter fragments too. The interface does not look very convincing. While the disordered region in MFF is likely to be a functional motif, the 4-helix bundle domain in PLP1 that AF models it to bind to is known to be a transmembrane domain, so the binding site is actually buried inside the membrane. AF is also not very confident about the domain structure, especially for the parts that are at the membrane surface or outside of it. The prediction is likely wrong.

run17: PAX6-CSNK2A1

CSNK2A is a widely active kinase, involved in many processes. Overlapping fragments from PAX6 show trend of increasing confidence the shorter the fragment. CSNK2A1 is predicted to bind with its kinase domain (it doesn't really has anything else than the kinase domain) to a peptide in PAX6 which seems to be a good looking linear motif, i.e. conserved, not part of a folded domain as predicted by AF and predicted by AF to form an alpha helix. The motif though overlaps with a putative NLS. The PAX6 motif is predicted to bind clearly to a pocket that exists in N-lobe of the kinase domain at the bottom of it, away from the catalytic side. Digging deeper, I found a structure, 1JWH, that shows that this is the pocket that is bound by CSNK2B, the regulatory subunit, that interacts with the catalytic subunit to form an active holoenzyme. This, however, does not eliminate the possibility that the AF prediction is right since the peptide looks like a functional motif.

run18: PAX6-SET

Top prediction is ordered-ordered, PAX6 Homeodomain and SET NAP domain. The structure 6PAX shows the PAX domain consisting of two similar folds like the homeodomain bound to DNA but the three-helix bundles are not oriented in exactly the same way like in the homeodomain so I am having a hard time to see where the homeodomain would bind DNA; AF models the homeodomain interface with the NAP domain of SAP via a charged interface with a lot of positively charged residues on the homeodomain contacting a patch of negatively charged residues on the SAP domain. It could be that this patch of positively charged residues on the homeodomain would usually interact with the negatively charged backbone of DNA, but the predicted structure from AF looks interesting since the interface likely does not interfere with SET homodimerization (2E50).

run19: PAX6-TLK2

All predictions with >0.7 model confidence are paired with the Pkinase domain of TLK2 and they are all predicted to bind at the bottom of the beta barrel fold (N-lobe) of the kinase domain. However, almost all peptides come from very different regions in PAX6, no recurrent predictions here.

When looking at the motif pLDDT metric then top predictions also involve two distinct motifs predicted to bind to the long helices in TLK2. However, AF predicts the two helices to form intramolecular contacts. By taking them apart into separate fragments it could be that intramolecular contact sites are now used for interface prediction.

The pair of interactions has a DMI predicted, MOD_GSK3_1 (PAX6 395-402). The peptide PAX6 394-404 was paired with the Pkinase domain but similar to the previous point, it is also put at the beta barrel fold in the N lobe and not the substrate binding site.

run20: PAX6-NGLY1

The PUB domain from Q96IV0, NGLY1, gives good model confidence, >0.8, in binding overlapping disordered fragments of P26367, PAX6. The PUB domain has been solved before alone (2CCQ), the catalytic domain has also been solved bound to RAD23 (2F4M); in the paper that published the PUB domain structure (Allen et al JBC 2006, 10.1074/jbc.M601173200) they also did some mutational analysis to show that there is an interface on the PUB domain that binds the AAA ATPase domain of p97 but the experimental evidence looks not very convincing. Indeed, AF modelled the peptide from PAX6 to bind to an interface adjacent to the one found by Allen et al. There is indeed some hydrophobic pocket and the best 4 predictions comprise that peptide binding to this pocket, however, which hydrophobic residue of the peptide is docked into the pocket varies depending on the length of the peptide; I think that this region in PAX6 could indeed be a linear motif, it is adjacent to the homeobox domain but I don't think that it is part of the homeobox domain.

run21: PAX6-ESRRG

Many short fragments with high model confidence that are scattered over the disordered region. The binding pocket on ESRRG is in the hormone receptor domain and is a known binding pocket for binding to L..LL motifs (ELMDB: LIG_NRBOX).

According to ELMDB, the first and last L go into a hydrophobic pocket and all fragments of PAX6 with high model confidence have more or less two hydrophobic amino acids with three residues in between: PAX6 319-329: DTA**L**TNT**Y**SA, PAX6 203-213: RLQ**L**KRK**L**QR, PAX6 374-384: PPH**M**QTH**M**NS, PAX6 198-208: DEAQ**M**RLQ**L**K, PAX6 128-148: GADG**M**YDK**L**R.

Looking at structures with ESRRG and two different bound peptides: 1KV6 and 1TFC: NCOA1 686-700: RHKI**L**HRL**L**QEGSPS, 2GPO and 2GPP: NRIP1 378-387: SL**L**LHL**L**KSQ, it furthermore became apparent that the hydrophobic residues right before both Leucines are also important for binding since they contact a hydrophobic patch on the other side of the pocket. However, none of the AlphaFold predicted motifs really fit, it is thus questionable whether they can actually bind the pocket.

Structurally speaking, the peptide does not fit that nicely in the hydrophobic pocket. In 2GPO and 2GPV, there is a triad of hydrophobic residues (V/L/I) making contact with the hydrophobic pocket on the domain but here only 2 residues are making contact. Therefore, it seems doubtful to me that this is a motif that can bind to the domain.

run22: PAX6-QRICH1

Difficult to dig deeper because QRICH1 has only one domain (DUF) that binds to C terminus peptide from PAX6. The high confidence peptide is 20 aa long and seems nice with 0.88 model confidence.

The same DUF is also modelled with 0.76 confidence with a very long disordered region (85 aa) that is at the N terminus of PAX6. However, the predicted complex of this disordered region is quite odd, as it has many twists and turns that seem weird to me.

Overall, these predictions look good but it's hard to be very certain about it because nothing is known about the domain in QRICH1 and PAX6 has a long disordered C-terminal region full of S, T, but also some Ps and hydrophobics.

run23: PAX6-KCTD7

The top prediction involves the disordered region of PAX6 (198-208) and BTB_2 domain of KCTD7, with 0.74 model confidence. No trend of increasing confidence when fragments shorten. InterPro describes this domain as one that multimerises for its protein function, e.g. KCTD1 as a transcriptional repressor (3DRX, solves KCTD5 that has a similar fold but shorter in length). Since BTB domain mediates the multimerisation of KCTD, it could be that it requires a certain stoichiometry for binding to its partner. In the HuRI database, KCTD7 was indeed detected to interact with itself. The two highest predicted models put both peptides into the same pocket and both peptides have some sequence similarity albeit from different regions in PAX6. These peptides were also predicted with high model confidence in other runs. Based on the structure 5FTA, BTB domains in their homodimerized form do expose the surface predicted in the top prediction. Therefore, the surface predicted to bind to the peptide would be available. Taken together, the prediction looks plausible.

run24: TTC19-FH

Not inspected because none of the predictions returns model confidence or average interface pLDDT above cutoff (model confidence ≥ 0.7, ordered-disordered prediction with disordered fragment interface plddt ≥ 70; ordered-ordered prediction with intf_avg_plddt ≥ 75).

run25: PEX3-PEX16

PEX3 and PEX16 are two proteins that seem to cooperate to help inserting new peroxisome membrane proteins (PMPs) into the peroxisome membrane. They do so via interaction between PEX3 and PEX19. PEX19 brings the PMPs to the peroxisome where PEX3 and PEX16 sit and mediate then further insertion of the cargo (this is described in review Smith and Aitchison 2013 Nature Rev Mol Cell Biol in Fig 2). However, there is also a study that describes how PEX16 localizes to ER and from there traffics to peroxisomes (Kim et al JCB 2006). The structure 6AJB that has been solved for the interaction between PEX3 and PEX19 was published by Sato et al EMBO J 2010 and describes how an N-terminal SLiM in PEX19 binds to the domain of PEX3. They tried to crystalize the whole protein of PEX3 but only observed residues 52-368. The domain has the exact same fold as predicted by AF. The predicted cytosolic and peroxisomal localization of protein regions and the two TM helices that are shown in Uniprot seem to be wrong for PEX3 according to work cited in Sato et al. They summarize that the N-terminal region of PEX3 contains a targeting signal or anchor for PEX3 to the peroxisomal membrane followed by the domain that is located in the cytosol. No structure has been solved yet for PEX16 but it seems likely that the prediction of two TM helices that are shown in UniProt in this protein is also wrong. AF predicts a globular domain containing the two TM helices and has a nicely exposed loop that carries the putative SLiM that AF predicted to bind to PEX3. It binds onto PEX3 on the opposite side to PEX19 binding, so PEX19 and PEX16 could bind simultaneously to PEX3. Further work on these interactions can be done by submitting the three protein sequences to AF to see what it does. Some other study observed interaction between PEX3 and PEX16 according to Uniprot but the interface really does not seem to have been looked at before nor the interaction studied in detail. All the fragments that contain the putative SLiM in PEX16 are predicted in the exact same way to bind to PEX3; always anchored via a conserved region sitting in PEX16 between residue 160 and 190. Interestingly, the most conserved residues are also those that seem most important for binding. This smells really good.

run26: PEX3-PEX19

This is a positive control interaction since the structure has been solved for this PPI (3AJB) and it is a well known and well studied PPI with an entry for it in the ELM DB: LIG_Pex3_1 (L..LL...L..F). This ELM instance is indeed predicted by AF to be the highest model confidence. Another peptide from PEX19 121-141, FTSCLKETLSGL, scored equally high model confidence. It could be that this other predicted binding site is also true but I believe that it is rather an artefact from AF's insensitivity to mutations.

run27: GABARAPL2-UBA5

The structure 6H8C shows binding of GABARAPL2 domain to LIR motif in UBA5. This motif is not listed on the ELM website for LIG_LIR_Gen_1 because it does not quite fit the regular expression which seems to be defined too narrowly. AlphaFold correctly predicts this interface but only as third highest based on model confidence just hitting the cutoff of 0.7 while using chainB_inf_avg_plddt it scores as fourth best prediction far below the cutoff (67). However, AF recurrently finds peptides including the motif following each other when ranked by model confidence or pLDDT. The top three motifs predicted to bind to GABARAPL2 are not finding the hydrophobic pocket that is filled by a key big hydrophobic residue in the motif and these peptides are also not recurrently predicted. So, I think these are wrong predictions.

run28: GABARAPL2-LZTR1

GABARAPL2 (P60520) has Atg8 domain that is known to bind motifs (LIG_LIR_Gen_1). The domain is modelled with high confidence to bind to different disordered fragments of interacting partner LZTR1 (Q8N653). The second top confident model (when ranked by model confidence) has an aromatic residue tucked into a deep pocket and a branched alipathic residue tucked into another shallow pocket. The top confident model has some kind of increasing trend in model confidence as fragments get shorter, with the shortest one getting the highest confidence. The highest confidence model has a nice increasing model confidence trend but it does not have an aromatic residue fitting into the deep pocket as it is known for LIG_LIR motifs.

Looking at the structure 2LUE, the second top model LZTR1 46-52 GP**F**ET**V**H looks more similar in sidechain positioning compared to 2LUE. Residues highlighted in bold get tucked into the mentioned pockets. This model seems more likely to be true than the best model. However, it also is predicted to bind in reverse order compared to structure 3WIM.

run29: CUL3-KCTD7

Has an ordered-ordered prediction with quite high confidence (0.66) but the contact interface is a tetramerization domain from KCTD7. Therefore it seems unlikely that it is a functional interface.

Two N terminus disordered fragments from KCTD7 with > 0.7 model confidence when paired with the Cullin domain of CUL3. These two fragments are modelled to be binding at the same site of Cullin domain (the site where RING proteins bind to, 1LDJ). In the case of 1LDJ, the RING protein has a long disordered region inserted into the Cullin domain of CUL1, burying a series of hydrophobic residues in the long disordered region. However, the same binding site of the Cullin domain of CUL3 is a bit different, with more surface exposed than CUL1. In this case, the contacts modelled in KCTD7 16-26, with a triple Serine making contact with the Cullin domain, look plausible. The other high confidence peptide KCTD7 1-11, with triple Valine making contact with the Cullin domain, also looks plausible to me.

In the structure of 1LDJ it is really amazing how the partner protein interacts with CUL1 via beta-sheet augmentation but how this extra beta strand becomes part of the integral fold, it is kind of in the middle of the domain. I think AlphaFold feels that there is something missing and is trying to put a peptide there but the overall conformation of the domain is also different at places so that the predicted peptide does not sit at the same position like the one shown in 1LDJ. AlphaFold predicts two different motifs of very different sequence from the N-terminus of KCTD7 to bind there. Given how different the sequences are, this adds another negative point towards questioning the specificity of these predictions.

run30: PNKP-SYP

Top prediction is a disordered fragment from SYP (7-19) paired with the kinase domain of PNKP. The binding surface is different from the nucleotide binding surface (1RC8). This binding interface looks plausible. It was later found that the kinase and phosphatase domain form a structural unit based on published structures. The run is modified to use the kinase and phosphatase domain as an ordered region for prediction with disordered fragments of SYP.

The rerun with a fragment comprising the phosphatase and kinase domain now resulted in one prediction that makes the cutoff. This prediction puts a motif from SYP into the DNA binding pocket of the kinase domain (according to Bernstein et al Mol Cell 2005, 1RC8).

There is another predicting docking a peptide from SYP into the FHA domain of PNKP. It puts it where FHA domains bind their phosphorylated peptides but the SYP peptide has no Ser or Thr.

run31: PNKP-TRIM37

The first prediction involving the combined kinase-phosphatase structure puts a peptide of TRIM37 into the binding pocket where the phosphatase domain would bind single stranded DNA.

Following up is a prediction that involves a disordered region in PNKP binding to the surface of MATH domain of TRIM37 where MATH domain-binding peptides generally bind to. The PNKP peptide differs slightly in sequence from regular expression patterns described for MATH domains in the ELM database. This peptide in PNKP has a known phosphorylation site that stabilizes PNKP protein levels, making the peptide very interesting since this suggests a regulatory role of phosphorylation on the peptide.

There is a second peptide of PNKP predicted to bind to the MATH domain also with high confidence but the sequence is quite different from the first one and very close to the phosphatase domain. There is also a prediction where the FHA domain of PNKP is predicted to bind to a peptide of TRIM37 but the peptide looks very different from known FHA-binding motifs (peptide with phosphorylated threonines), which is of course difficult to predict for AF.

run32: PNKP-XRCC4

XRCC4 and PNKP prediction, there is a peptide from XRCC4 that binds to the phosphatase domain with high confidence. But then I am not sure if this is right because it could be a false prediction of a small peptide easily fitting into the catalytic site of the phosphatase domain. There is a Serine in the peptide, so it is possible that this is where the phosphate group gets cleaved off by the phosphatase domain. After checking more, it is found that XRCC4 is known to bind to PNKP via a phosphorylated motif that binds to the FHA domain in PNKP.

In principle, it would be better to make a rerun where the kinase and phosphatase domain are taken as one fragment since they form 1 structural unit but I think in this case it would not have changed anything. The best prediction put a peptide from XRCC4 into the

pocket of the phosphatase domain where it would bind the single-stranded DNA as seen in 3U7G. Among the first 9 predictions AF put 7 different peptides from XRCC4 into the phosphatase, the others go to the kinase domain. The first prediction that involves the FHA domain of PNKP and contains the FHA-binding motif in the sequence fragment of XRCC4 has a confidence score of 60 and does not put the FHA-binding motif in the pocket but another negatively charged peptide in the sequence (the FHA pocket is very positively charged). The correct prediction where AF puts the FHA-binding motif in the right pocket has a confidence score of 0.58.

run33: TNPO3-GCH1

Top prediction involves the disordered region of GCH1 (16-26) and the superhelical structure of TNPO3, with model confidence 0.71. Since TNPO3 (transportin) is known to transport cargo into the nucleus by releasing the cargo via the competitive binding of GTP-bound Ran (2X19), the peptides from GCH1 are modelled to be at a binding site near where Ran binds in 2X19. It is therefore biologically sound where the peptides are modelled at. The binding site of the peptides from GCH1 is also lined with many arginines, making it very positively charged. The contact modelled by AF in the top prediction looks good, with many charge-charge interactions at the interface. The N terminus of GCH1 has many prolines that are conserved, with three repeats of PAEK or PEAK and two repeats of PPRP.

run34: TNPO3-CAMK2G

Not inspected because none of the predictions returns model confidence or average interface pLDDT above cutoff (model confidence ≥ 0.7, ordered-disordered prediction with disordered fragment interface plddt ≥ 70; ordered-ordered prediction with intf_avg_plddt ≥ 75).

run35: GNAI3-GPSM2

This interaction has been structurally solved (4G5S) and AlphaFold predicted the interface 100% accurately. GPSM2 has multiple GoLoco motifs that AlphaFold predicts individually with high confidence to bind in the pocket on GNAI3.

run36: SYT1-MIP

Both are transmembrane proteins. The top prediction involves the linker between two C2 domains of SYT1 and the MIP domain of MIP. MIP domain is also known as aquaporin domain (transmembrane). However, when the linker is fragmented, it receives lower confidence. I think this is unlikely to be the interaction interface. The linker could be a motif for some other interaction because of its moderately high plddt. There is a structure of a homodimer of SYT1, 2R83, that shows that both C2 domains of one chain are actually interacting with each other and that the linker between both domains interacts with one domain. It is this linker where AF predicts that a peptide would bind to the porin domain of MIP; interestingly, AF predicts the two C2 domains to be independent from each other in the monomeric structure of SYT1, so either AF is wrong or crystallization introduced the packing of both domains against each other but I would rather believe the Xray structure and in this case the peptide would not be accessible to bind to the MIP domain.

run37: FTSJ1-CERT1

FTSJ domain of FTSJ1 is known to bind S-ADENOSYLMETHIONINE (see structure 1EJ0). The top predictions all look very different in that different regions or partially overlapping regions of CERT1 are docked into different sites onto the FTSJ domain. Sometimes the

peptide is docked into the catalytic pocket where the protein methylates adenosines on tRNAs but the peptide is also docked elsewhere. Because of these ambiguities, I believe that the predictions are questionable since they seem to lack specificity but I don't think we can call them definitely wrong.

Another interface was found involving CERT1 368-388, with model confidence 0.70. However, the contacts modelled are mostly backbone-to-backbone. I have previously noticed that AF tends to give higher confidence to complex modelled with secondary structure. So I think this is also a likely false interface.

run38: CAMK2A-SOX5

Kinase domain of CAMK2A with a disordered fragment predicted showed high confidence. The structure predicted by the highest confidence model is weird, with both beta sheet and helix structure.

Kinase domain of CAMK2A is likely serine threonine kinase and in kinase domain prediction, one has to be careful with the two lobes that bind substrate and ATP. It might be interesting to check other high scoring peptide to see if they have S/T that can be phosphorylated and check the crystal structure to find substrate binding pocket. The first two highest scoring peptides do not look convincing because the first one has no S/T in the peptide but it is fit into the catalytic cleft while the second one has positioned the sidechain of a T out of the cleft. The third highest scoring peptide (P35711 131-141) looks nice because it positions the sidechain of an S into the catalytic cleft.

The highest ranking peptides are essentially all over the place from SOX5 and I don't think that AF can predict very well kinase-substrate interactions. Overall, the high-scoring predictions all do not look very convincing.

run39: CAMK2A-CAMK2G

Many high confidence predictions involving different regions in the protein pair. Among them, one ordered-ordered interface gives a really high confidence. The interface is a known DDI in 3did with high zscore. The structure 3SOA only shows one CAMK2A monomer but the publication talks about a dodecamer for which one can download a model from the PDB as well. Looking at this dodecamer and the paper, it becomes clear that downstream of the kinase domain there is another domain referred to as hub domain in the paper which mediates oligomerization, together with the linker between the kinase and hub domain. The best AF prediction for the interaction between CAMK2A and CAMK2G involves both hub domains and is an accurate prediction of the interface seen in the dodecamer.

The second best prediction made by AF involves the hub domain and a bit of the linker sequence from the other partner. Looking at the dodecamer, one can see that where the peptide is predicted to bind on the hub domain is part of the linker sequence bound from the same monomer, so an intra-molecular interaction. So, there is indeed some binding site but not for inter-molecular interaction. Because the linker sequences are different in the structure and canonical uniprot sequence it is very difficult to know which part of the linker is binding on the hub domain and whether this corresponds to the bit of the linker sequence predicted by AF to bind there. In the paper accompanying the 3SOA structure they also investigate how different linker sequences from different isoforms influence Ca-binding site accessibility and thus activation of the complex. There is evidence from 3 other studies that CAMK2G and CAMK2A interact with each other from co-IP experiments but these were large-scale studies. It is likely that no one has studied the interface between CAMK2G and CAMK2A and thus would be something new.

run40: ACTB-ACTG1

Two actin proteins are predicted to have high confidence DDI. The interface itself that is predicted by AlphaFold looks very interesting, it indeed looks like a polymerization interface because both domains interact with opposite sites. interactome3D would model this interaction with the structure 4JHD as a template but this one looks quite different, it's not the same interface and needs according to the authors a third protein for polymerization. Digging deeper in PDB for structures of ACTB, I found structure 6ANU which shows the same interface that AF predicted between ACTB and ACTG1, so the interface is probably right.

This is also a very interesting case. Based on the review by Vedula and Kashina (J of Cell Signal 2018, 10.1242/jcs.215509), it is still an open question whether the different actin forms that exist in human can form heteropolymers or not. Some studies find this in vitro, other find intermingled homopolymers of beta and gamma actin. Both actins co-occur in many cell types while alpha-actin is more specifically expressed in muscle. It seems really tricky to solve this since actins are highly studied and actins are also super similar in their sequence, so it could be that in a somewhat artificial system, beta and gamma actin can interact because the interface residues are identical but in vivo they would rather not interact and rather form homopolymers. In the end, whether ACTB and ACTG1 indeed interact in vivo is the only open question.

run41: RARB-PSMC5

PSMC5 has been repeatedly modelled by AF to have a high confidence peptide that binds to partners with Hormone_recep domain. The peptide is 132-141 DP**L**VS**LM**MVE. Residues highlighted in bold are the ones tucked into the hydrophobic pocket. However, this peptide does not match with the consensus of LIG_NRBox (^PL..LL^P), especially in this peptide P precedes the first L. I am not sure why P is disallowed at first position as ELM has not described much about the sequence composition of the motif. I think it might be too early to reject this peptide because the highlighted residues are indeed hydrophobic and can serve similar functions as those in the regex.

I looked at the HuRI network of PSMC5 too, and found that the interactors seem to be enriched with the Hormone_recep domain, making this interface even more plausible.

run42: DCX-BICD2

DCX has two DCX domains and all good predictions involve the N terminus DCX domain. The N terminus DCX domain is known to bind Tubulin. AF modelled a different interface on the N terminus DCX domain to bind to disordered fragments from BICD2.

The DCX domains have a C-terminal part that is not confidently predicted by AF to be part of the fold. When excluding this part from the first DCX domain, AF models peptides to bind to the area where this last part is predicted to be located in the monomeric structure from AF. When we use a DCX domain that contains this last bit, then AF predicts other peptides from BICD2 to bind on the opposite side of DCX. There is no consistency in these predictions.

There are no other predictions between ordered-ordered or disordered fragments binding to ordered domains in BICD2 that make the cutoffs. BICD2 however, also only consists of large helices. Nonetheless, it could be that both DCX domains together bind to one of these coiled coil helices in BICD2.

run43: DCX-ZBTB10

A possible prediction involves the first DCX domain of DCX and a peptide of ZBTB10 261-271. This prediction is not influenced by the actual domain boundaries because the peptide is not docked into the pocket where a region a little C-terminal of the domain might bind to. This is the case for the second best prediction involving the first DCX domain and peptide 604-614. According to chainB inf avg plddt these are the only two prediction that make the cutoff when looking at chainB as a disordered region. ZBTB10 has a lot of disorder and probably many motifs. DCX has two DCX domains and a bit of disorder. Looking into available PDB structures then the DCX domains are known to bind to microtubules. There is one structure with the first DCX domain bound to microtubules (6RFD). It seems though that the pocket where ZBTB10 261-271 is predicted to bind is not occupied in this complex. AF does not predict slightly extended versions of this peptide with reasonable confidence to bind to this pocket.

A peptide was also predicted to bind in beta-sheet augmentation to the last beta strand of the BTB domain with reasonable model confidence and chainA_intf_avg_plddt scores but the ZBTB10 model might have its own beta strand C-terminal of the current domain boundaries that AF predicted to complement the last beta strand of the domain as predicted in the full length model of ZBTB10.

AF also predicts a contact between the ZnF domain of ZBTB10 and the first DCX domain but it does not look very likely and I think the ZnF fold is perturbed.

run44: PSMC5-ESRRG

The interaction has quite some high confidence predictions. The highest scoring peptide is P62195, PSMC5, 132-141, DP**L**VS**LM**MVE. The three hydrophobic residues make nice contacts with the hydrophobic pocket and surface of the domain. Another disordered fragment from PSMC5 binding to the same domain, IKKLWK, also looks promising. However, there is some possibility that these are artefacts because AF is not very specific when it comes to detecting single mutation in known motifs. The sequence alignments are not helpful unfortunately because the whole PSMC5 is super conserved.

Nonetheless, interaction between PSMC5 and ESRRG looks promising because the alternative name is thyroid hormone receptor-interacting protein 1, TRIP1.

run45: PSMC5-RORB

The highest confidence prediction involves a disordered fragment from PSMC5 and it is the same as run44. The ordered region from RORB is the same domain, hormone receptor domain, as run44.

It is interesting to see AF predicting similar DMI with high confidence from two different proteins. Same observation as run44.

run46: WAC-NFE2L2

WAC and NFE2L2 are largely disordered. WAC has a WW domain. AF predicts recurrently a sequence close to the N-terminus of NFE2L2 to bind to the WW domain that are known to bind proline-rich motifs. The putative motif in NFE2L2 does not contain prolines and is not docked onto the WW domain in any way like other WW domains, e.g. 1EG4. These are likely wrong predictions. While the motif interface pLDDT is reasonably high for these predictions, the model confidence does not reach the 0.6. There are no other predictions that make the cutoff.

run47: WAC-MOBP

Not inspected because none of the predictions returns model confidence or average interface pLDDT above cutoff (model confidence ≥ 0.7, ordered-disordered prediction with disordered fragment interface plddt ≥ 70; ordered-ordered prediction with intf_avg_plddt ≥ 75).

run48: STX1B-FBXO28

The top model has 0.76 model confidence that utilizes the disordered region 1-22 of STX1B and Fbox + helix bundle domain (63-221) of FBXO28. The interface involves the disordered region of STX1B forming a 310 helix structure with the helices from the Fbox domain. Note that the Fbox domain annotated by InterPro is from 61-109, while the ordered region that I used for prediction is 63-221. The Fbox domain is known to mediate PPI but it is not used by AF to model the interaction in this prediction. Region 1-22 of STX1B is conserved only in recent homologs. The plddt of the disordered region is low, <60 for all residues.

The second top model has 0.75 model confidence that involves the syntaxin domain (23-237) of STX1B and disordered region 354-368 of FBXO28. The disordered region of FBXO28 is at the C terminus and conserved. However, the plddt of the peptide is low and adopts a 310 helix kind of structure. A slightly different prediction involving fragments of the proteins (27-219 STX1B and 345-363 FBXO28) returned 0.73 model confidence. The peptide adopts a helical structure but is placed on a different surface of the syntaxin domain. Although the peptide 345-363 has good plddt (mosty >60), I am not sure if this is the right interface. One prediction pairs the full length of STX1B with the disordered region 354-368 of FBXO28 and returned 0.71 model confidence. The interface is similar to that of the syntaxin domain (23-237) of STX1B and disordered region 354-368 of FBXO28 with low plddt. This region 354-368 in FBXO28 could be an nuclear localization signal (NLS), where ELMDB also predicts quite a few NLS, and therefore unlikely to be the interface for the interaction.

Next top prediction has 0.749 model confidence that involves the C terminus of the syntaxin domain (220-232) of STX1B as disordered region and the Fbox + helices domain of FBXO28 (63-221). The interface is formed by the peptide adopting a helical structure with the Fbox + helices domain. The plddt of the peptide is good, with all residues above 60 plddt. Nonetheless, another prediction involving slightly longer peptide from the same region of STX1B has a much lower model confidence (0.55). The interface modelled is not exactly the same as it is a little bit shifted. Unsure if this is a good interface.

I tried to find more molecular studies on the two proteins but I can't find much. STX1B is known to function in docking of synaptic vesicles at presynaptic active zones while FBXO28 probably recognizes and binds to some phosphorylated proteins and promotes their ubiquitination and degradation. Weirdly, STX1B is known to localize to membrane while FBXO28 has not much information on subcellular localization but studies have shown that it interacts with topoisomerase using its Fbox domain (the bundled helices are not needed for interaction). Out of all the predictions, I think STX1B 27-219 + FBXO28 345-363 and STX1B 220-232 + FBXO28 63-221 are most likely to be the interface, as their peptides are modelled with good plddt and both achieved model confidence higher than 0.7.

run49: STX1B-MMGT1

Top prediction involves the Syntaxin domain of STX1B and the disordered region of MMGT1 (23-31) with confidence 0.73. A slightly longer fragment has a slightly lower confidence but looking at the structure, the two peptides have different angles to the Syntaxin helical bundle. Since the interfaces modelled by AF differ a lot despite using the same peptide and its extended counterpart, the modelled interfaces do not look genuine.

run50: STX1B-VAMP2

Interactome3D models an interface between both proteins based on the structure 3HD7/3IDP where STX1A interacts with VAMP2. STX1A and STX1B are very similar in structure.

STX1B is predicted in closed conformation, which we know because structures exist of STX1A bound to Munc18 where it is in this closed conformation with the long C-terminal helix comprising the SNARE domain folding back onto the syntaxin domain. However, when bound to VAMP2 we can see the open conformation where the long helix is made available to bind in coiled-coil like manner to VAMP2 and SNAP25 helices.

Based on this available structural information we designed different fragments of the extended SNARE domain of variable length. VAMP2 is a short protein of 116 residues consisting of a long helix and about 30 disordered residues at the N-terminus. The most confident predictions obtained for these fragments is the one modeling a coiled-coil interaction between the extended SNARE domain and the helix of VAMP2 but the model confidence is slightly below the cutoff. Predictions with the disordered N-terminal region of VAMP2 remain far below cutoffs.

run51: CSNK2A1-CSNK2B

Nice prediction with overlapping fragments showing increasing model confidence. This interface has been solved before in two structures: 4DGL and 6Q38; prediction is highly accurate, and is probably a DMI that is not in ELM yet.

run52: EBF3-EBF2

Dimerization of the EBF family already known and solved (3MUJ). AF predicts the middle domain of both proteins called TIG as the dimerization interface as top prediction but in head to tail orientation while the structure 3MUJ shows head to head orientation. Followed closely up in terms of score (avg_intf_plddt) is the fragment comprising the TIG domain and the helix loop helix domain which are predicted accurately as seen in the structure.

The third best prediction involves the N-terminal DNA binding domain as the dimerization interface. Does not look so convincing to me but still got a very high score. The fourth best prediction is the helix-loop-helix domain alone as dimerization interface, still with a score of 90. There are more predictions that make the cutoff that involve various disordered regions of either protein and ordered fragments from the other involving interfaces used for dimerization but I guess that these predictions are likely wrong.

run53: PEX12-TREX1

The disordered region of PEX12 215-312 (98 residues long) is predicted with high confidence. One fragment of it achieved even higher confidence but when this fragment is further fragmentate, their confidence is not as high anymore. After checking the protein on InterPro, this domain is the exonuclease domain of TREX1 that binds to ssDNA (2OA8). In this crystal structure, it shows the pocket modelled by AF to bind PEX12 215-312 is bound to a ssDNA, with the phosphodiester bond of ssDNA making interactions with the backbone of the domain chain and some hydrophobic side chain (leucine) making hydrophobic interaction with the base of the nucleotide. Interestingly, AF seems to have memorized this crystal structure because the bound ssDNA has a curved structure and AF also models the long disordered region to have an odd curve. I think this interface is unlikely to be true because the bound magnesium ions coordinate with the oxygen in the phosphodiester bond of ssDNA and the modelled helix places hydrophobic sidechains to the cavity where magnesium ions bind.

A very short fragment of PEX12 12-16 at the N terminus is modelled with high confidence with a very negatively charged pocket in the domain of TREX1. It is unusual to have a peptide binding pocket with such a high negative charge. Further checking revealed that this domain binds magnesium ion and nucleotides. The short fragment fits into the magnesium binding pocket and thus this is unlikely to be true.

run54: PRKAR1A-PRKAR1B

Best model is an ordered-ordered prediction with 0.83 confidence. It is a homo-DDI (RIIa domain) dimerization and has been solved in 2EZW.

An additional disordered fragment (PRKAR1A 360-372) predicted with high model confidence but low pLDDT with the cyclic nucleotide binding domain of PRKAR1B. Referencing available structure of cNMP binding domain (1NE4), there are two beta barrel folds in the domain that bind to cyclic nucleotides. AF fits the disordered fragment on a hydrophobic surface near the beta barrel but not in the cNMP binding pocket. Although this could be another binding site, the binding makes little sense to me because the disordered fragment is at the C terminus of cNMP binding domain of PRKAR1A, meaning that the sequence would have to loop back to make this contact. In the previous bullet point, it seems very likely that the dimerization of the two proteins are mediated by the RIIa domain (N terminus), so it seems not so plausible to me that at the C terminus they make contact again. This is likely a false positive interface.

run55: ASF1A-H4C8

The interaction between both proteins has been solved (5C3I). However, this structure shows that the motif in H4 sits at the very C-terminus and binds in beta sheet augmentation to ASF1A in the same pocket like AF predicted but using an N-terminal peptide of H4. I think the problem is that the C-terminal region of H4 was made part of the domain of H4, which I agree was hard to see from looking at the monomeric AF structure for full length H4; I checked further down in the predicted structures but the first ordered-ordered prediction has a model confidence of 0.25 and does not find this mode of binding either. One could rerun this by taking the C-terminal peptide of H4 as disordered region just to see whether AF would then get it right but in principle this is a false positive prediction; the N-terminal peptide also shares no sequence similarity with the C-terminal motif.

run56: RARS1-CCDC115

There is only one prediction that makes the cutoff for model confidence or/and motif pLDDT. This prediction involves RARS1 1-21 as a disordered fragment that is modelled to bind as a helix to the two helix coiled-coil domain of CCDC115. A shorter fragment of the motif is placed elsewhere. The helix of CCDC115 to which the peptide is predicted to bind has more hydrophobic residues along the helix on that side so I would think that a longer partner chain would be able to bind there. Thus, this interface does not seem likely to be true.

run57: UBE3A-TAT

Not inspected because none of the predictions returns model confidence or average interface pLDDT above cutoff (model confidence ≥ 0.7, ordered-disordered prediction with disordered fragment interface plddt ≥ 70; ordered-ordered prediction with intf_avg_plddt ≥ 75).

run58: VAMP4-MFF

Top prediction is two ordered regions that are both helical. Both proteins have only helical regions and the rest are disordered. Interestingly, despite the top predicted interface having only 0.71 model confidence, both chains have very high plddt for their residues at the interface (95 for VAMP4 and 90 for MFF). Because of their high plddt, it could be a genuine interface. The helix in VAMP4 definitely has an interface there because one side is rather hydrophobic while the other side is rather hydrophilic. MFF could bind there with its helix or via another helix that it has. The binding does not show that many nice contacts, i.e. some hydrophobic residues on the VAMP4 helix still remain exposed.

run59: PEX16-MMGT1

Not inspected because none of the predictions returns model confidence or average interface pLDDT above cutoff (model confidence ≥ 0.7, ordered-disordered prediction with disordered fragment interface plddt ≥ 70; ordered-ordered prediction with intf_avg_plddt ≥ 75).

run60: PLP1-SLC16A2

Not inspected because none of the predictions returns model confidence or average interface pLDDT above cutoff (model confidence ≥ 0.7, ordered-disordered prediction with disordered fragment interface plddt ≥ 70; ordered-ordered prediction with intf_avg_plddt ≥ 75).

run62: SNRPB-GIGYF1

GIGYF1 is a very long protein with many disordered regions. It has a GYF domain that is known to bind proline-rich sequences. SNRPB has many proline-rich sequences in its C terminus. Some proline-rich motifs are predicted with high pLDDT to bind the GYF domain (these are the top predictions).

Another highly ranked prediction involves the LSM domain of SNRPB with various disordered fragments from GIGYF1. However, checking InterPro entry as well as structures showing LSM domain, it seems like LSM domain is predominantly involved in multimerization with other SNRP proteins to form the SMN complex involved in splicing (1H64). Therefore, the models involving this domain with disordered fragments look unlikely to be true to me.

Digging deeper into the top predictions, comparing the binding modelled by AF between SNRPB 231-240 and GYF domain of GIGYF1 with 1L2Z, the peptide is oriented differently. However, from 3FMA, one can see different ways a peptide binds to the same surface of GYF. In 3FMA chain E and P show a similar way of binding to that modelled by AF. The peptide sequence in 3FMA is also different from 1L2Z, but importantly, there are three prolines in the peptide that always orient the same to the hydrophobic surface formed by the GYF motif on the GYF domain. This orientation of the 3 prolines is captured by AF.

AlphaFold repeatedly predicts the PPGM motif in the same pocket. This motif occurs multiple times in the C-ter tail of SNRPB. On the ELM website, the LIG_GYF motif is described to bind proline-rich sequences and they also cite the structure 1L2Z but they say that flanking positively charged residues seem to be important for binding to the GYF domain. Indeed, in the crystal structure there are some negatively charged residues on the GYF domain. Interestingly, the GYF domain from GIGYF1 does not or only partially has those. It also differs in that it has a deeper hydrophobic pocket which is filled with a Trp in the crystal structure. So, it could well be that the GYF domain from GIGYF1 binds somewhat different proline-rich peptides. The interaction between GIGYF1 and SNRPB has not been described before other than in HuRI. Functionally, it would be probably a new connection because GIGYF1 is not known to function in splicing as far as I can see and thought to be localized to the cytoplasm. GIGYF1 however, has also interacted with SNRPA and SNRPC in HuRI. They also have 1 or

some more occurrences of the PPGM motif. If this mode of binding is true then it would be somewhat of a new mode of binding or in the most conservative case an extension of the known binding mode of LIG_GYF.

Alignment of 1L2Z chain A (GYF domain) with the GYF domain from GIGYF1 (476-535) shows that the sequences are not very conserved. Structural superimposition of the two GYF domains reveal that the overall fold is conserved, including the majority of the binding pocket except for the hydrophobic pocket filled with a W. The peptides of the two structures have their PPPG in similar orientation. Following this sequence is a M from SNRPB that is tucked into the hydrophobic pocket and H for 1L2Z that is exposed to the environment. The sequence that follows is R for both, with the one in SNRPB exposed to the environment and possibly forming a hydrogen bond with the Q on the domain, and that in CD2 (1L2Z) forming salt bridge with an E from the domain.

Later a structure of the GYF domain of GIGYF1 was published binding to a similar motif found in TNRC6 further supporting the correctness of these predictions.

run63: ARHGEF9-VEZF1

Top prediction has 0.74 model confidence with the fragment from VEZF1 (375-385) making contact with the RhoGEF domain of ARHGEF9. The top predictions all put the peptide at the same binding site of the RhoGEF domain. In terms of conservation, all the peptides from VEZF1 are well conserved. Nonetheless, the prediction looks like a very questionable one, at least it seems like the predictions do not make use of the GTP/GDP binding pocket for which I did not find a structure that shows where it precisely is located but based on an abstract of an article and InterPro entries it seems to be between both structural entities that form one larger domain, the GEF domain and the PH domain (IPR000219). There is absolutely no consistency in the two peptides from VEZF1 selected to bind to the same surface on the GEF domain of ARHGEF9; VEZF1 also seems to be of very weird type, AF has a hard time to make sense out of this protein.

run64: MIP-MFF

Not inspected because none of the predictions returns model confidence or average interface pLDDT above cutoff (model confidence ≥ 0.7, ordered-disordered prediction with disordered fragment interface plddt ≥ 70; ordered-ordered prediction with intf_avg_plddt ≥ 75).

run65: VEZF1-PRKAR1B

Not inspected because none of the predictions returns model confidence or average interface pLDDT above cutoff (model confidence ≥ 0.7, ordered-disordered prediction with disordered fragment interface plddt ≥ 70; ordered-ordered prediction with intf_avg_plddt ≥ 75).

run66: VEZF1-KCTD7

Top prediction involves the disordered region of VEZF1 (360-380) and the BTB domain of KCTD7. The disordered region overlaps with the top prediction in run63 that models the interface between VEZF1 (375-385) and the RhoGEF domain of ARHGEF9. Despite AF modelling a 310 helix structure in the disordered region of VEZF1 (360-380), the contacts modelled at the interface do not look very convincing. It could be that the disordered region (360-385) is a functional motif for other interactions and AF detects that and tries to fit it into the domain. It could also be that, to form the binding interface, it needs multiple copies of BTB domain, which is not used in this prediction. The VEZF1 peptide is put in the same pocket like

the PAX6 peptides from run23 but the sequences look different, it is however the same peptide in VEZF1 like in the prediction with ARHGEF9.

run67: APTX-FLAD1

Has overlapping fragments with increasing confidence: APTX, N terminus disordered region 5-12 and 6-13, paired with MoCF_biosynth or a domain of unknown type (not matched to a Pfam or SMART domain) that is between the MoCF and PAPS_reduct domain of FLAD1. It also predicts the same N-terminal region of APTX into the PAPS_reduct domain. The disordered fragments from the region 8-15 of APTX showed high confidence model confidence but below the cutoff pLDDT score when modelled with the PAPS_reduct domain of FLAD1. Checking the structure of PAPS_reduct domain in complex with adenosine phosphosulfate shows that the peptide is modelled by AF to be in the binding pocket of adenosine phosphosulfate. This is likely a false prediction.

For the N-terminal part of APTX AF is quite confident when it models it into the MoCF domain or the other unknown domain of FLAD1. There are multiple predictions with different overlapping fragments that make the cutoff. However, AF is more confident with both metrics when the peptide is modelled into the MoCF domain. This domain has a pretty substantial pocket that is actually in the monomeric structure of FLAD1 occupied by another region of FLAD1 with low pLDDT. However, when APTX 10-15 is used for modelling, the orientation of the peptide is reversed. MoCF_biosynth domain is known to trimerize for its activity and is known to bind molybdopterin. MoCF_biosynth binds molybdopterin on a site close to where AF models the peptide to be (refer to 1DI6, https://doi.org/10.1074/jbc.275.3.1814 that solves the structure of a bacterial protein with the same domain. They mentioned 49D and 82D to be important for catalytic activity)

APTX with the unknown domain of FLAD1 does not reach the model confidence cutoff, only the motif pLDDT cutoff. It puts the same peptide as beta-sheet augmentation to the domain while in the predictions for the MoCF domain, the peptide is put in helical conformation.

The only predictions where disordered regions in FLAD1 are predicted to bind to folded regions in APTX involve the FHA domain of APTX and correspond to two completely different disordered regions in FLAD1.

run68: FBXO28-PSMC3

Top prediction is coiled-coil interaction between regions from the two proteins that are modelled by AF monomer as long helices. The plddt of all residues are very high. This interaction looks convincing. The only problem is that one helix is shorter than the other, while for a common coiled coil interaction, both helices are usually equally long.

The second best prediction based on model confidence involves a disordered region from FBXO28 (51-61). The modelled complex does not look convincing because the peptide is quite hydrophobic and the residues do not make much contact with the domain. The peptide is predicted to bind to the first domain of PSMC3 which as far as I was able to find, does not have catalytic activity.

There are only these two predictions that make the cutoff for model confidence, none make the cutoff when looking for disordered regions in PSMC3 predicted to bind to FBXO28. The other way round there is the peptide mentioned above and a C-terminal disordered region of FBXO28 predicted to bind to the same first domain in PSMC3 but predicted to bind to a different side. The C-terminus of FBXO28 is very charged, maybe a localization signal. Both motifs in FBXO28 are somewhat recurrently predicted to bind to the domain in PSMC3.

run69: CAMK2G-ESRRG

Many high confidence predictions in a disordered region of CAMK2G. The whole disordered region used as a fragment for prediction also returned high confidence (0.78). In this long disordered region, AF puts the third highest model confidence peptide in the domain pocket. The top three highest confidences are very similar in terms of confidence. The motif detected by AF resembles LIG_NRBOX with the motif L..LL. CAMK2G 300-310: LKGAI**L**TT**ML**V -> looks plausible to me because the M is hydrophobic and it is possible to substitute for the role of L in the regex. CAMK2G 315-325: SA**A**KS**LL**NKKS -> Also possible but the A is fitted into a quite deep hydrophobic pocket where known structure (refer to run21) shows that it is L that gets fit into the pocket. A might have too short of a hydrophobic side chain to make a good contact with the deep pocket. CAMK2G 355-365: QEPAP**L**QTAME -> not so good IMO because the hydrophobic contact is less extensive as the peptide found above. Another interesting observation: CAMK2G 285-423 (139 aa) prediction resulted in 0.78 model confidence, which is very high for a disordered region that long. In this case, **CAMK2G 300-310** is fitted into the hydrophobic pocket, adding weight to the fact that this could be the correct peptide. This reminds me of the extension analysis with DMI where extension of motif can improve prediction results.

A pairing of ordered-ordered region prediction returned high confidence (0.83). This involves Zn finger from ESRRG and CaMKII association domain at the C terminus of CAMK2G. The binding is close to but not in the Zn binding pocket, which is good. CaMKII association domain of CAMK2 has been shown to oligomerize with other CAMK2 in 1HKX.

Looking at the monomeric structure of ESRRG and CAMK2G, it looks possible that the C terminus association domain of CAMK2G to bind to ESRRG via Zn finger domain of ESRRG and the hormone receptor domain of ESRRG binds to the long and disordered region separating the two domains found in CAMK2G. This makes a multi-site binding between two proteins and a very interesting case.

run70: XRCC4-LIG4

The structure for this interaction has been solved: 3II6 and 1IK9. Looking at the structure of 3II6, the two proteins interact with each other via XRCC4 first forming a homodimer with its coiled-coil domain, then around the homodimer binds the tandem BRCT domains of LIG4. The BRCT domains are separated by a structurally less defined region that most likely forms two helices upon binding to XRCC4. Not sure if this can be seen as domain-motif or domain-domain interaction, probably something in between. It is not so clear from the monomeric AF model of full length LIG4 that both BRCT domains form a functional unit but I guess one could have also made a fragment comprising both domains and the linker sequence. Runs so far were made with both BRCT domains individually and the linker sequence individually and further rerun has to be done by using the BRCT domain tandem as one structural unit.

The top prediction involves a motif at the C-terminus of XRCC4 that is predicted to bind to the last BRCT domain of LIG4. I think the prediction is wrong because of the solved structure. The prediction also does not look like how other motifs bind to BRCT, i.e. the protein FANCJ (LIG_BRCT_BRCA_1). However, the C-terminus of XRCC4 certainly carries one or two motifs. One is annotated in Proviz as WD40 domain binding. The very C-terminus is a class 3 PDZ-binding motif. The whole region is very conserved. Maybe this is why AF tries to put peptides from this C-terminus in various domains, including the DNA ligase domain of LIG4 (fourth top prediction). So, the top two predictions involve this C-terminus and reach high confidences in both metrics (model confidence and intf_avg_plddt).

The third highest prediction involves the XRCC4 N-terminal domain plus one long helix (taken as one ordered region) and the 2nd BRCT domain. This interface is exactly the same interface that is seen in the structure 3II6 where part of the BRCT domain also contacts the XRCC4 helix.

The 6th best prediction involves the linker between both BRCT domains and the XRCC4 helix. Despite the fact that XRCC4 is in monomeric form in our prediction and that the BRCT domains are missing, AF correctly models the contacts between the linker and the single XRCC4 domain as they can be seen in the structure 3II6. This model meets both cutoffs, for model confidence and pLDDT.

Rerun using the BRCT domain tandem as one structural unit completed. The tandem BRCT fragment ranks 7th with the coiled coil XRCC4 fragment based on model confidence and second for ordered-ordered fragment pairs when ranked by avg interface plddt. The prediction that is still ranked first is the single BRCT domain binding to the coiled coil fragment (92 vs 89 avg intf plddt score).

run71: TMEM237-MFF

Not inspected because none of the predictions returns model confidence or average interface pLDDT above cutoff (model confidence ≥ 0.7, ordered-disordered prediction with disordered fragment interface plddt ≥ 70; ordered-ordered prediction with intf_avg_plddt ≥ 75).

run72: HNRNPK-TH

In the full length structural model of HNRNPK the first 2 KH domains are predicted to pack against each other using an interface that is also predicted to bind to the TH peptide 61-71. This region indeed overlaps with a Pfam HMM that seems to find some pattern in this disordered region but nothing is known about this “structural”(?) motif. It predicts 3 occurrences of it in the N-terminal region of TH but the third one is the most conserved and this is the one predicted to bind to the second KH domain. Two other motifs overlapping with 61-71 are also predicted to bind to this KH domain. The residues that are part of all three motifs are predicted to bind to the KH domain in the same way. One prediction below the model confidence cutoff predicts the motif to bind to the third KH domain but in a different way.

run73: OTX2-RPS26

Not inspected because none of the predictions returns model confidence or average interface pLDDT above cutoff (model confidence ≥ 0.7, ordered-disordered prediction with disordered fragment interface plddt ≥ 70; ordered-ordered prediction with intf_avg_plddt ≥ 75).

run74: MFF-MMGT1

Not inspected because none of the predictions returns model confidence or average interface pLDDT above cutoff (model confidence ≥ 0.7, ordered-disordered prediction with disordered fragment interface plddt ≥ 70; ordered-ordered prediction with intf_avg_plddt ≥ 75).

run75: PUF60-TH

The top prediction involves using both RRM domains of PUF60 as one ordered region and a disordered polyA peptide from TH. The peptide is put at the same position where the Nbox would bind as shown in the NMR structure 2KXH. However, the predicted peptide has some different sequence: solved structure: LxxAxxI, model: VxxAxxV, and there are no recurrent predictions. Another prediction involves the third RRM domain of PUF60 and another peptide in TH which tugs a Trp in a pocket but it does not look very convincing.

Prediction involving disordered fragments from PUF60 and ordered region (Biopterin_H domain) from TH returned a maximum of 0.78 model confidence. This is likely false interface because the short peptide is fit into the biopterin and iron binding pocket of the enzymatic domain (refer to run72 for example). The second best prediction is also fitted at the same site, therefore also likely a false interface.

Interestingly, the disordered region of PUF60 302-461 is modelled with 0.69 model confidence with the Biopterin_H domain of TH. The long disordered region makes contacts with two regions of the domain, one at the iron binding site (likely false) and another coiled-coil interaction at the C terminal helix of Biopterin_H domain. This coiled-coil interaction is repeated in a shorter disordered fragment of PUF60 (317-347, third best prediction (0.77), the same C terminal helix in the long disordered region). This coiled-coil interaction looks like a plausible interface.

I tried finding more information about this ACT-like domain but to no avail. InterPro says that it homo-dimerizes using the beta strands like in 1Q5V, but the fold is not exactly the same. The ACT-like domain in TH is special in the way that the last beta strand is formed by its N and C termini by looping back to meet each other. I cannot find much information about this domain.

run76: PUF60-QRICH1

One long disordered region of PUF60 (1-128) is modelled with high model confidence with DUF of QRICH1. In this region, 111-121 is modelled at the interface. This region when fragmented from the long disordered region also showed high confidence (0.86). This fragment tucks a R into a very deep negatively charged pocket but the rest of the peptide seems to make questionable contact with the DUF domain.

Top prediction with ordered region in QRICH1 and peptides in PUF60 either put the linker helix between the first two RRM domains or the N-terminal long helix in PUF60 or another helical peptide at 442-461 at two different places on the DUF domain. I think that the helical linker between both RRM domains is not accessible for this mode of binding because the key residues are making intramolecular contacts to the RRM domains in the AF monomer PUF60 model.

3 different peptides are predicted to bind to the tandem PUF60 domain. In principle, the long disordered N-terminal region of QRICH1 is full of potential helical peptides of pattern hydrophobic-x-x-Ala-x-x-hydrophobic, which is the kind of peptide that is like the Nbox motif that can bind to PUF60 and the three different peptides are also predicted to bind to the same pocket.

There are also 4 different peptides in QRICH1 predicted to bind to the third RRM domain.

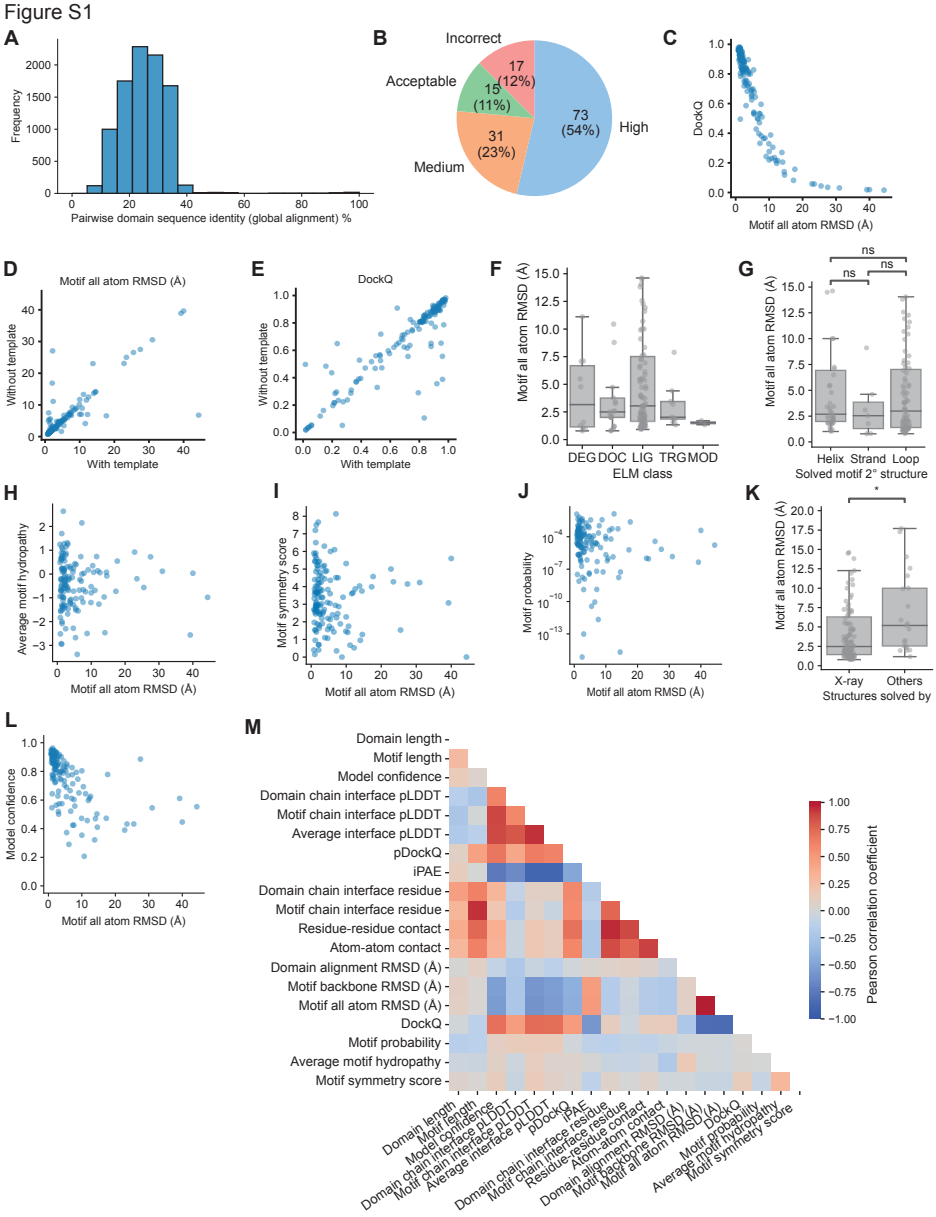
run77: MAB21L2-AP1S2

The top prediction involves Clat_adaptor_s domain of AP1S2 with the disordered fragment (215-220) of MAB21L2 (78 motif pLDDT, 0.77 model confidence). The motif is predicted recurrently with variable length but the disordered region is generally very short because it is a loop within the domain of MAB21L2. AF also made a disulfide bridge between motif and domain. Not sure this is correct. Looking at the structure 1W63 that shows the large Ap1 clathrin adaptor core complex where there is a fold similar to the one in AP1S2, one can see that the region where the peptide is predicted to bind would in principle be accessible for binding. This domain Clat_adaptor_s is known to bind motifs from ELMDB but no structure has been solved in terms of this domain and its bound peptide. The disordered fragments from

the previous point also do not match with any ELM class that binds to Clat_adaptor_s. Other good predictions use the Mab-21 domain of MAB21L2. Two overlapping disordered fragments (146-154, 0.68 and 153-157, 0.75) had good confidence with the domain but they are modelled to be at different binding sites, so it does not look likely to me that this is the binding region.

run78: PRKAR1B-QRICH1

The motif in PRKAR1B is at the very C-terminus of the protein and also matches a PDZ-binding motif. There is only one prediction that makes the model confidence cutoff but it does not meet the pLDDT cutoff. The C-terminal peptide of PRKAR1B binds to the only domain of QRICH1 but extended or smaller versions of the motif are only predicted with very low score then to bind to the domain so no recurrence here. The prediction therefore looks unlikely to be functional. No other predictions make the pLDDT cutoff.

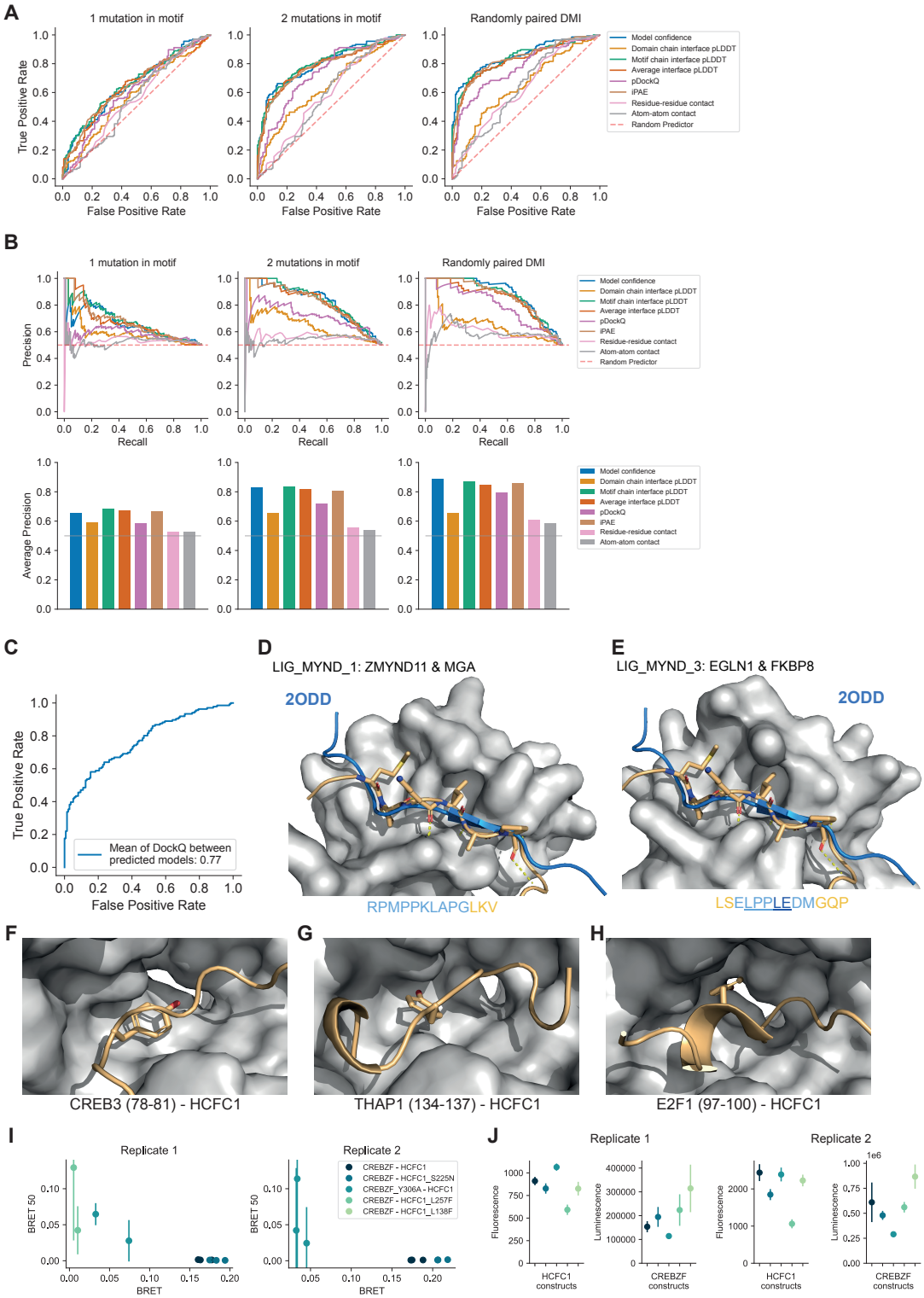


Appendix Figure S1. Benchmarking of AF on DMI interfaces using minimal interacting regions.

A Pairwise sequence identity of domains in the DMI positive reference dataset. **B** Proportion of high, medium, acceptable and incorrect models predicted by AF from the positive reference dataset as classified by the DockQ score. **C** Scatterplot of DockQ vs motif RMSD for DMIs from positive benchmark dataset. Pearson r = -0.85, p-value < 0.0001. **D-E** Motif RMSD and DockQ scores of structures for DMIs from positive benchmark dataset predicted by AF with and without the use of templates. Motif RMSD: Pearson r = 0.81, p-value < 0.0001. DockQ: Pearson r = 0.88, p-value < 0.0001. **F** Accuracy of AF DMI predictions stratified according to the annotated functional categories of DMIs in the ELM DB. DEG=degron, DOC=docking, LIG=ligand, TRG=targeting, MOD=modification. **G** Accuracy of AF DMI predictions stratified according to the secondary structure element formed by the motif in the solved structure. **H-J** Scatterplot of various motif features vs motif RMSD determined for models and structures of DMIs from positive benchmark dataset: H motif hydropathy, Pearson r = -0.03, p-value = 0.72, I motif symmetry, Pearson r = -0.08, p-value

= 0.38, J motif regular expression degeneracy, Pearson r = -0.04, p-value = 0.66. **K** Accuracy of AF DMI predictions stratified according to the method used to solve the structures in the benchmark dataset, Mann-Whitney-Wilcoxon test two-sided p-value = 0.017 test statistics = 811 **L** Scatterplot of model confidence of predicted models vs motif RMSD determined from superimposing the predicted models with structures of DMIs from the positive benchmark dataset. Pearson r = -0.55, p-value < 0.0001. **M** Correlation matrix of different prediction variables and prediction outcomes.

Figure S2

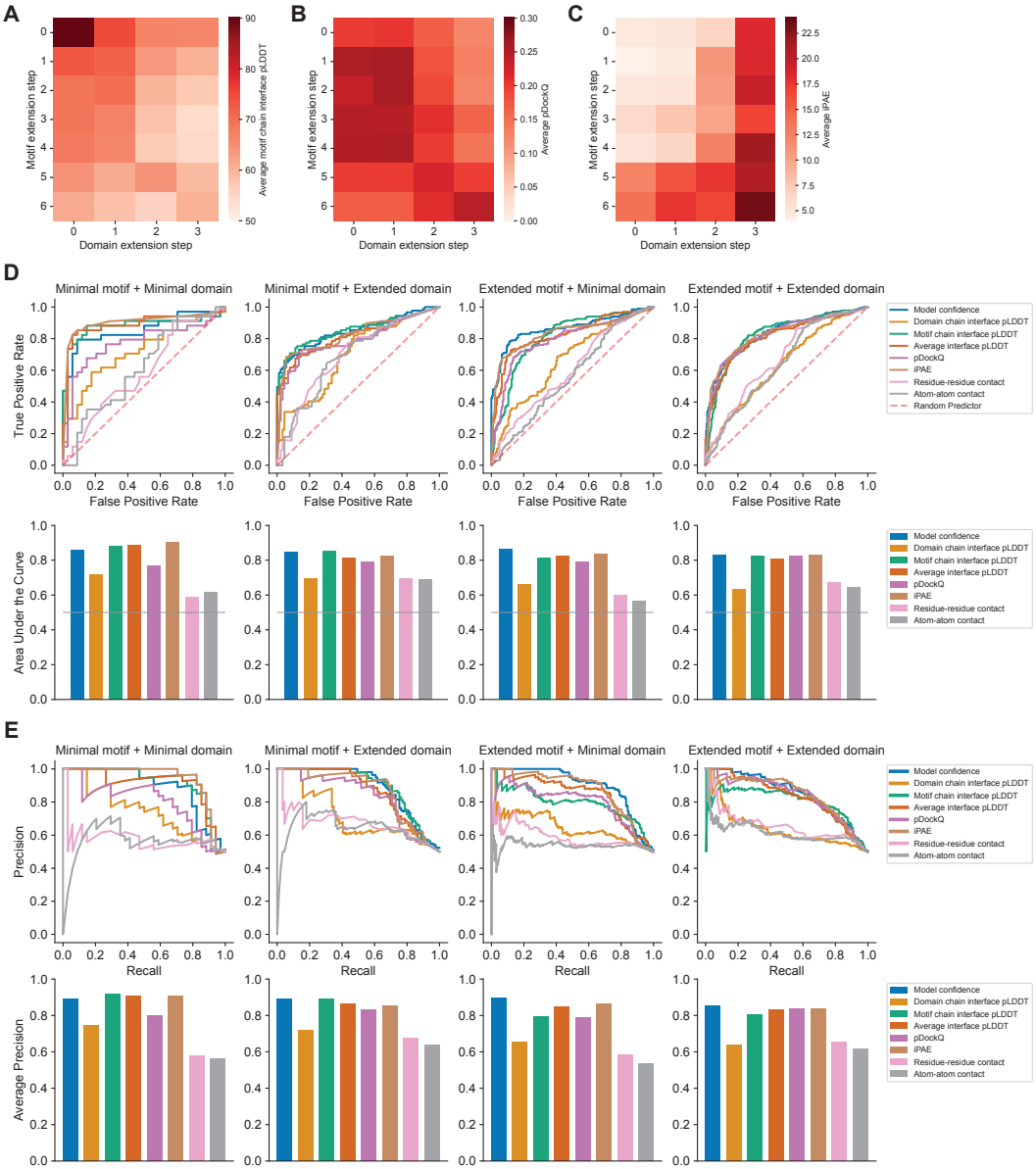


Appendix Figure S2. Benchmarking and application of AF for DMI interface prediction using minimal interacting fragments.

A Receiver operating characteristic (ROC) curve of various metrics extracted from AF models when using the DMI benchmark dataset as the positive reference and the following

sets as random reference: Left, 1 mutation introduced in conserved motif position; middle, 2 mutations introduced in conserved motif positions, right, randomly shuffled domain-motif pairs. **B** Precision recall curve of various metrics determined for benchmark datasets as in A. **C** ROC curve of mean DockQ between the top five AF structural models returned for a given input, assessed using the DMI positive reference set and random pairings of domains and motifs as in A. The AUROC of the metric is indicated in the legend of the ROC curve. **D-E** Superimposition of AF structural model for motif class LIG_MYND_1 (D) and LIG_MYND_3 (E) (orange) with homologous solved structures (PDB:2ODD) from motif class LIG_MYND_2 (blue). The motif sequence used for prediction is indicated at the bottom, colored by pLDDT (dark blue=highest pLDDT). **F-H** AF models for three motif instances (orange) of LIG_HCF-1_HBM_1 predicted to bind into a pocket on the Kelch domain of HCFC1 (gray). Motif positions are indicated below the figures. The key tyrosines of the motif sequences are drawn as sticks. **I** BRET50 estimates from fitting titration curves shown in Fig 1G are plotted vs. BRET values that were corrected for bleedthrough and measured at a 2:50 ng DNA transfection ratio for wildtype and mutant CREBZF-HCFC1 pairs. Error bars indicate the standard error. Data is shown for two technical replicates for the first biological replicate and three technical replicates for the second biological replicate. **J** Fluorescence and total luminescence are shown for wildtype and mutant CREBZF-HCFC1 pairs measured at a 2:50 ng DNA transfection ratio. Error bars indicate STD of two technical replicates for the first biological replicate and three technical replicates for the second biological replicate. Coloring as in I.

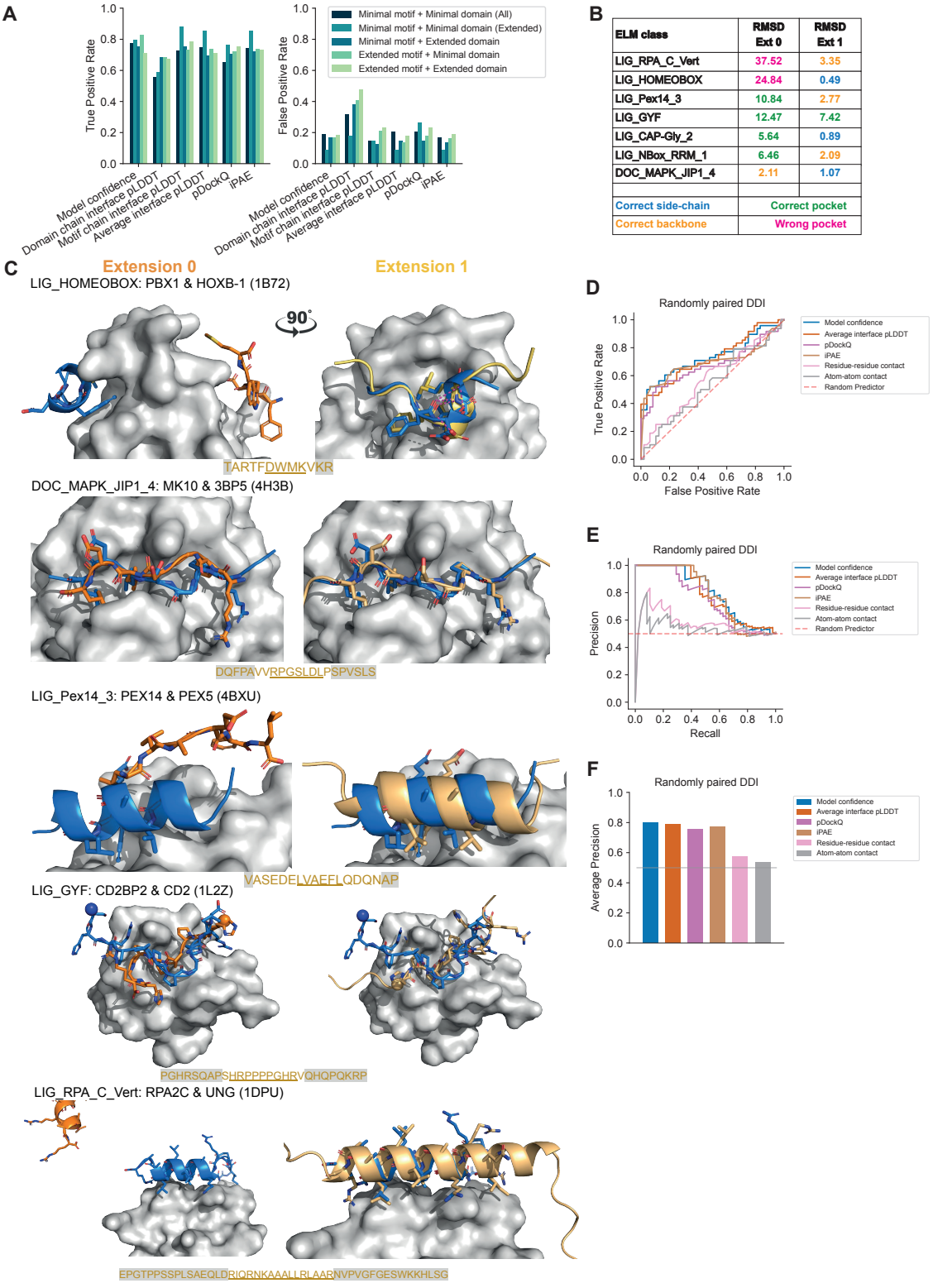
Figure S3



Appendix Figure S3. Effect of protein fragment extensions on the accuracy of AF predictions.

A-C Heatmap of the average motif interface pLDDT (A), pDockQ (B), and iPAE (C) for combinations of different motif and domain sequence extensions using a positive reference set consisting of 31 DMI structures. Extensions like in Fig 2A. **D** ROC curves (top) and corresponding AUROC values (bottom) of various metrics extracted from AF models when using the DMI extension dataset split by different combinations of motif and domain extensions as indicated on the top of each graph. Gray horizontal line indicates the AUROC of a random predictor. **E** Precision recall curves (top) and area under the precision recall curve as quantified by average precision (bottom) for various metrics extracted from AF models determined for benchmark datasets as in D.

Figure S4



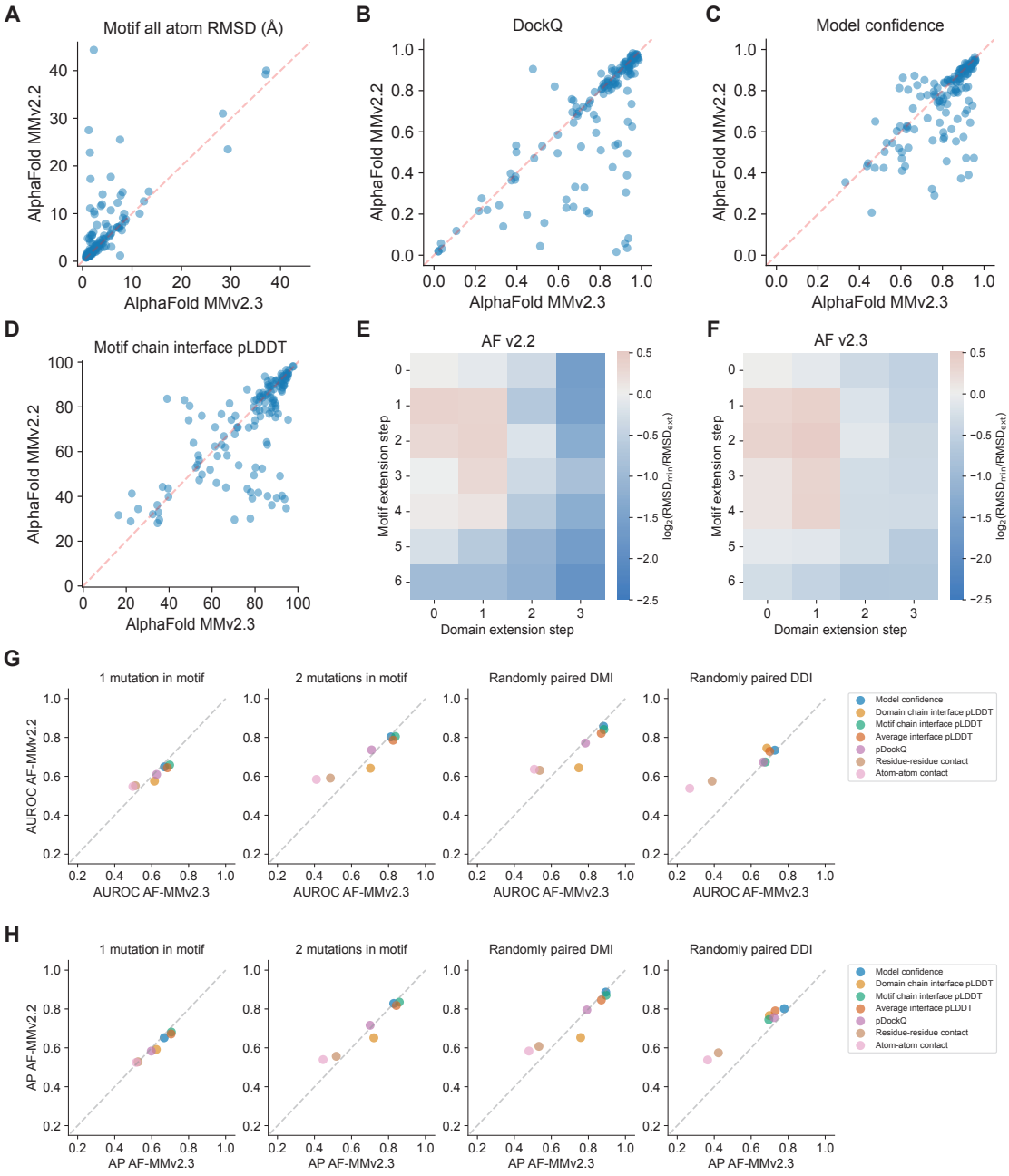
ELM class	RMSD Ext 0	RMSD Ext 1
LIG_RPA_C_Vert	37.52	3.35
LIG_HOMEOBOX	24.84	0.49
LIG_Pex14_3	10.84	2.77
LIG_GYF	12.47	7.42
LIG_CAP-Gly_2	5.64	0.89
LIG_NBox_RRM_1	6.46	2.09
DOC_MAPK_JIP1_4	2.11	1.07
Correct side-chain	Correct pocket	
Correct backbone	Wrong pocket	

Appendix Figure S4. Effect of protein fragment extensions on the accuracy of AF predictions.

A True and false positive rate (left and right, respectively) based on optimal cutoffs from Fig 2D derived for different metrics from ROC analysis for benchmarking AF with different motif

and domain extensions from the reference dataset illustrated in Fig 2A and random pairings of domain and motif sequences. **B** Table indicating the motif RMSD achieved when using minimal (extension 0) or extended motif sequences for structure prediction for all inspected motif extension cases. Extension 1 refers to extension of the minimal motif sequence by the length of the motif to the left and right. Color coding indicates the accuracy classes of the respective structural models as shown in Fig 1A. **C** Superimposition of the structural model of the minimal (left, orange) or extended (right, yellow) motif sequence with the solved structure (motif in blue) for five different motif classes as indicated on the top of each panel. The motif sequence from the solved structure is indicated at the bottom of each panel. Motif residues are underlined, motif residues not resolved in the structure have a gray background. Sticks indicate the motif residues, domain surfaces are shown in gray based on experimental structures. **D** ROC curves of different metrics using the DDI benchmark dataset as positive reference and random shuffling of domain-domain pairs as negative reference. **E** Precision recall curves of different metrics extracted from AF models determined for benchmark datasets as in D. **F** Area under the precision recall curve as quantified by average precision for metrics extracted from AF models determined for benchmark datasets as in D. Gray horizontal line indicates the average precision of a random predictor.

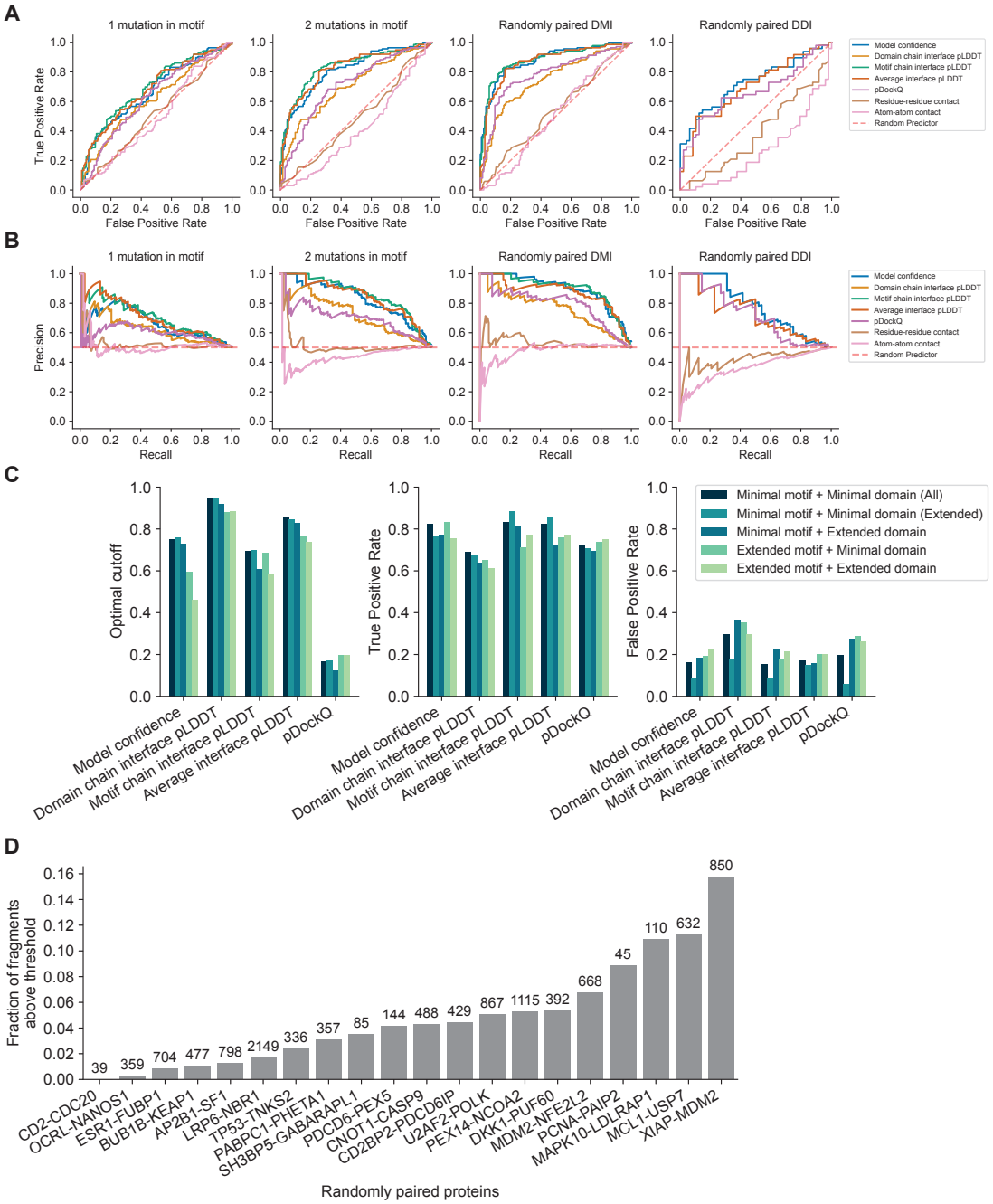
Figure S5



Appendix Figure S5. Comparison of AF v2.2 and v2.3 prediction performance. **A** Scatterplot showing the motif RMSD obtained from structural models computed either with AF v2.2 or AF v2.3 using the minimal interacting regions of all annotated DMIs. **B-D** Scatterplots computed as in A showing the DockQ (B), model confidence (C), and motif chain interface pLDDT (D) for both AF versions. **E-F** Heatmaps showing the fold change in motif RMSD obtained for structural models from AF v2.2 (E) and AF v2.3 (F) upon domain or/and motif sequence extension compared to when using minimal interacting regions. Positive values indicate improved predictions from extension and negative values indicate worse prediction outcomes. **G** Scatterplots showing the AUROC obtained for different metrics derived from structural models from benchmarking AF v2.2 and AF v2.3 using the minimal interacting regions of all annotated DMIs or DDIs as the positive reference dataset and different random reference datasets: Left (DMI), 1 mutation introduced in conserved

motif position; middle-left (DMI), 2 mutations introduced in conserved motif positions, middle-right (DMI), randomly shuffled domain-motif pairs; right (DDI), randomly shuffled domain-domain pairs. Corresponding ROC curves for AF v2.2 and AF v2.3 are shown in Fig. S2A, S4D, and S6A. **H** Scatterplots as in G plotting the average precision (AP) obtained from PR curves from the same analysis as in G. Corresponding PR curves for AF v2.2 and AF v2.3 are shown in Fig S2B, S4E and S6B.

Figure S6

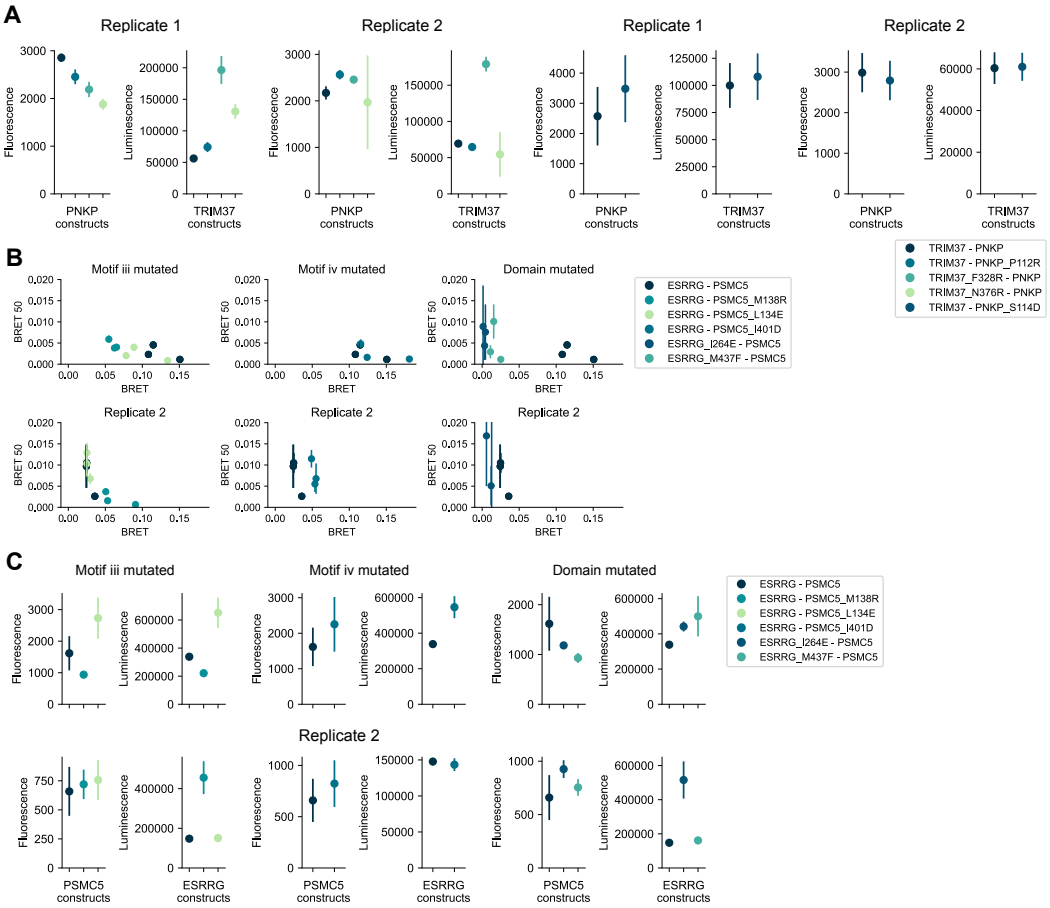


Appendix Figure S6. Performance of different metrics derived from structural models when benchmarking AF v2.3 for DMI predictions.

A ROC curves obtained for different metrics derived from structural models from benchmarking AF v2.3 using the minimal interacting regions of all annotated DMIs or DDIs as the positive reference dataset and different random reference datasets: Left (DMI), 1 mutation introduced in conserved motif position; middle-left (DMI), 2 mutations introduced in conserved motif positions, middle-right (DMI), randomly shuffled domain-motif pairs; right (DDI), randomly shuffled domain-domain pairs. **B** PR curves computed for the same datasets and AF version as in A. **C** Optimal cutoff, true, and false positive rate derived for different metrics from ROC analysis for benchmarking AF v2.3 with different motif and domain extensions from the reference dataset used in Fig 2A and randomly shuffled domain

-motif pairs. **D** Fraction of fragment pairs with structural models scoring above thresholds for 20 randomly shuffled domain-motif pairs. Numbers on top of the bars indicate the total number of fragment pairs submitted for interface prediction to AF for each random protein pair.

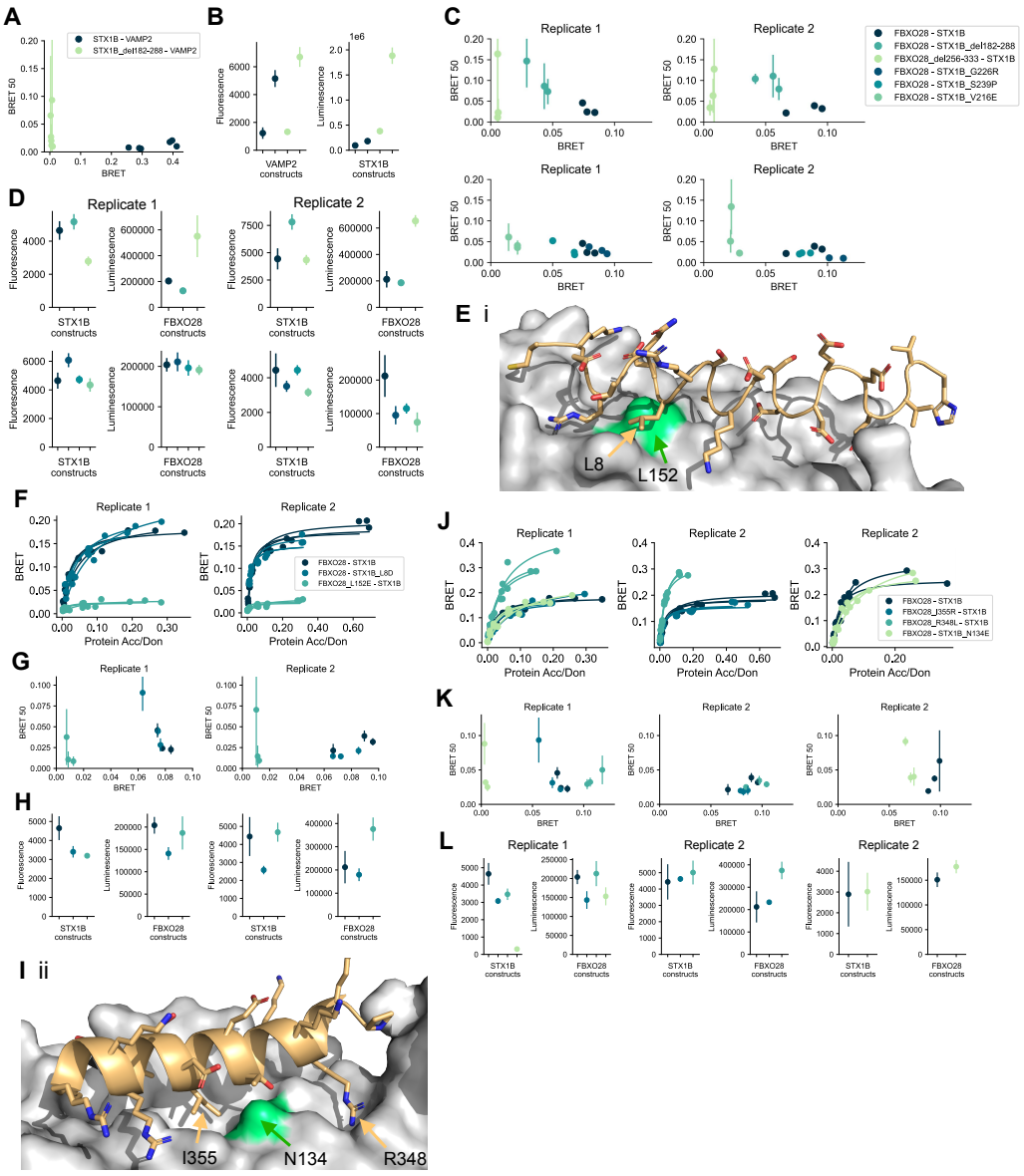
Figure S7



Appendix Figure S7. Expression and BRET50 plots for TRIM37-PNKP and ESRRG-PSMC5.

A Fluorescence and total luminescence are shown for wildtype and mutant TRIM37-PNKP pairs measured at a 2:50 ng DNA transfection ratio. Error bars indicate STD of three technical replicates. Data is shown for two biological replicates. **B** BRET50 estimates from fitting titration curves shown in Fig 4H are plotted vs. BRET values that were corrected for bleedthrough and measured at a 2:50 ng DNA transfection ratio for wildtype and mutant ESRRG-PSMC5 pairs. Error bars indicate the standard error. Data is shown for three technical replicates for two biological replicates each. BRET50 estimates for the second biological replicate for the ESRRG_M437F-PSMC5 pair were omitted from the graph because they exceeded the upper y-axis limit. Roman labels refer to interfaces shown in Fig 4E. **C** Fluorescence and total luminescence are shown for wildtype and mutant ESRRG-PSMC5 pairs measured at a 2:50 ng DNA transfection ratio. Error bars indicate STD of three technical replicates. Data is shown for two biological replicates.

Figure S8

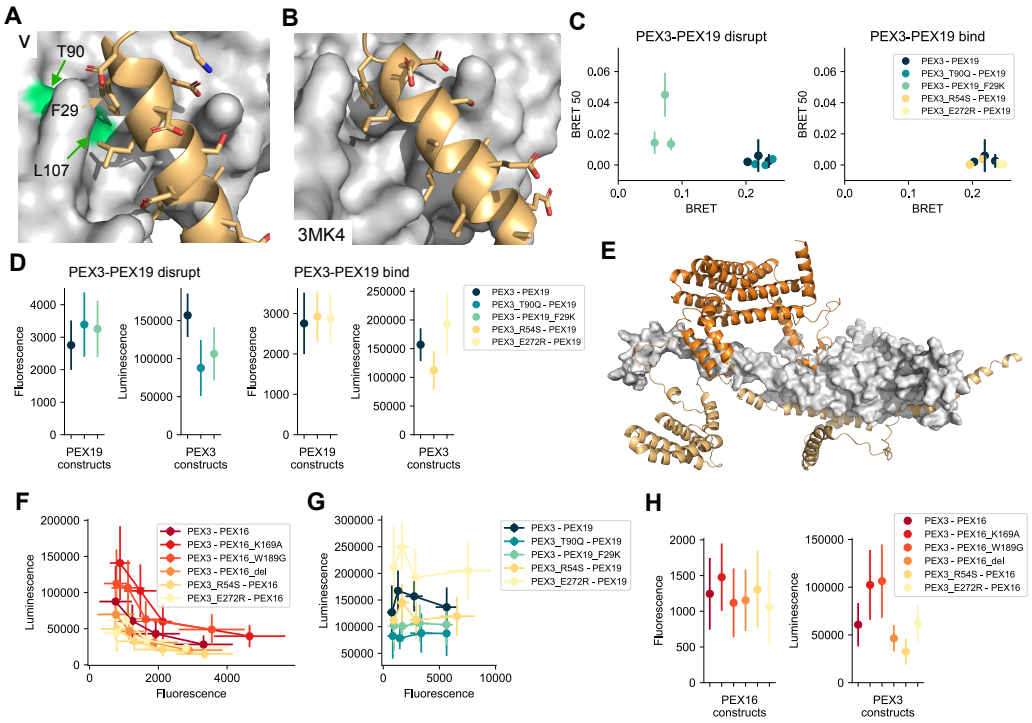


Appendix Figure S8. Structural models, expression, and BRET50 plots for STX1B-FBXO28 and STX1B-VAMP2.

A BRET50 estimates from fitting titration curves shown in Fig 5C are plotted vs. BRET values that were corrected for bleedthrough and measured at a 2:50 ng DNA transfection ratio for wildtype and mutant STX1B-VAMP2 pairs. Error bars indicate the standard error. Data is shown for three technical replicates for two biological replicates each. **B** Fluorescence and total luminescence are shown for wildtype and mutant STX1B-VAMP2 pairs measured at a 2:50 ng DNA transfection ratio. Error bars indicate STD of three technical replicates. Data is shown for two biological replicates. **C** Data shown as in A for wildtype and mutant FBXO28-STX1B pairs relating to interface iii (Fig 5A,D). **D** Data shown as in B for wildtype and mutant FBXO28-STX1B pairs shown in C. **E** Structural model corresponding to interface i shown in Fig 5A. Mutated residues on the domain (green) and motif side are labeled. **F** BRET titration curves are shown for wildtype and mutant FBXO28-STX1B pairs relating to interface i shown in E with two biological replicates, each with three

technical replicates. Protein acceptor over protein donor expression levels are plotted on the x-axis determined from fluorescence and luminescence measurements, respectively. **G** Data shown as in A for wildtype and mutant FBXO28-STX1B pairs relating to interface i. **H** Data shown as in B for wildtype and mutant FBXO28-STX1B pairs relating to interface i. **I** Structural model corresponding to interface ii shown in Fig 5A. Mutated residues on the domain (green) and motif side are labeled. **J** Data shown as in F for wildtype and mutant FBXO28-STX1B pairs relating to interface ii. **K** Data shown as in A for wildtype and mutant FBXO28-STX1B pairs relating to interface i. **L** Data shown as in B for wildtype and mutant FBXO28-STX1B pairs relating to interface i.

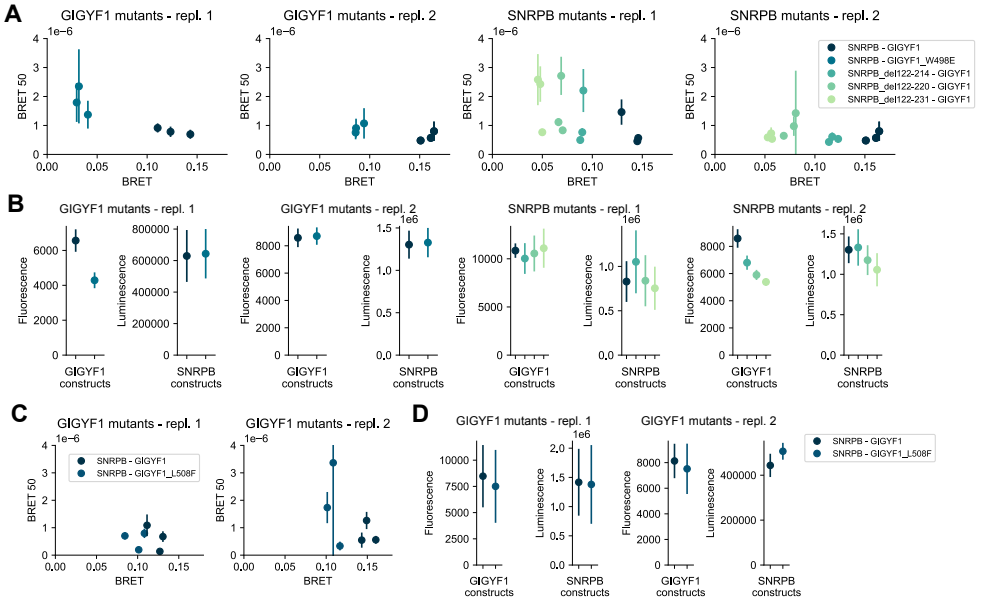
Figure S9



Appendix Figure S9. Structural models, expression, and BRET50 plots for PEX3-PEX19 and PEX3-PEX16.

A Structural model of PEX3-PEX19 corresponding to interface v as shown in Fig 5G. Mutated residues on the domain (green) and motif side are labeled. **B** Structure from PDB:3MK4 showing the PEX19 N-terminal motif bound to the PEX3 domain. **C** BRET50 estimates from fitting titration curves shown in Fig 5H are plotted vs. BRET values that were corrected for bleedthrough and measured at a 2:50 ng (for PEX3 and PEX3_T90Q) or 8:50 ng (for PEX3, PEX3_R54S, PEX3_E272R) DNA transfection ratio for wildtype and mutant PEX3-PEX19 pairs. Error bars indicate the standard error. Data is shown for three technical replicates. The left panel corresponds to mutant constructs that should disrupt binding while mutants shown in the right panel were aimed to disrupt binding to PEX16 and thus should not disrupt binding to PEX19. **D** Fluorescence and total luminescence are shown for wildtype and mutant PEX3-PEX19 pairs measured at a 2:50 or 8:50 ng DNA transfection ratio (see panel C). Error bars indicate STD of three technical replicates. **E** Structural model obtained with AF for the trimeric complex of PEX3 (gray), PEX19 (yellow), and PEX16 (orange) using full length sequences as input. **F** PEX3 expression levels measured in luminescence units plotted for co-transfections with increasing PEX16 protein amounts measured in fluorescence units. Error bars indicate STD of three technical replicates. **G** PEX3 expression levels measured in luminescence units plotted for co-transfections with increasing PEX19 protein amounts measured in fluorescence units. Error bars indicate STD of three technical replicates. **H** Data shown as in D for wildtype and mutant constructs of PEX3-PEX16 pairs. Measures are taken for 2:25 ng DNA transfection ratios.

Figure S10



Appendix Figure S10. Expression and BRET50 plots for SNRPB-GIGYF1.
A BRET50 estimates from fitting titration curves shown in Fig 6D are plotted vs. BRET values that were corrected for bleedthrough and measured at a 2:50 ng DNA transfection ratio for wildtype and mutant SNRPB-GIGYF1 pairs. Error bars indicate the standard error. Data is shown for three technical replicates for two biological replicates each. **B** Fluorescence and total luminescence are shown for wildtype and mutant SNRPB-GIGYF1 pairs measured at a 2:50 ng DNA transfection ratio. Error bars indicate STD of three technical replicates. Data is shown for two biological replicates. Coloring as in A. **C** Data shown as in A for wildtype and mutant SNRPB-GIGYF1 pairs fitted from titration curves shown in Fig 6E. **D** Data shown as in B for wildtype and mutant SNRPB-GIGYF1 pairs shown in C.

Chapter 3

Systematic domain-motif interaction interface and variant characterization using protein interaction profiling

3.1 Development of domain-motif interface predictor tool

To address the lack of mechanistic information on PPIs and the limitation of the current bioinformatic tool in the prediction of PPI interfaces, our former PhD student designed the DMI predictor tool. Here I will discuss the workflow of the tool, its performance and its application on HuRI interactome.

3.1.1 The workflow of the DMI predictor

The pipeline employed the UniProt identifiers for a pair of interacting proteins (e.g. A & B). Within these protein sequences, it uses Hidden Markov Models (HMMs) to identify the presence of known motif-binding domains. At the same time, regular expressions are applied to detect the occurrence of known motifs. Using a list of DMI types from the ELM database, the pipeline pairs the identified domains and motifs to generate putative DMI matches (**Figure 3.1 A**). These DMI matches are then annotated with features such as ANCHOR and IUPred scores (the propensity of motif disorderliness and the tendency to undergo a secondary structure upon binding with a partner), RLC score (motif conservation score across orthologs), the degeneracy of motif types based on their regular expression, the enrichment of the binding domain in the interaction partners and frequency of motif-binding domains (**Figure 3.1 B**). The matches are then scored using a Random Forest (RF) model.

To train and evaluate this model for predicting DMIs, a positive reference set (PRS) and several versions of a random reference set (RRS) were generated. The PRS is based on the 830 known DMI instances from the ELM database, while RRS was created by randomly pairing proteins and scanning for DMI occurrences (**Figure 3.1 B**). Each RRS version was paired up with the PRS to train separated RF models, and the performance was evaluated on test sets. Among these models, version 4 generated by randomly sampling DMI instances from the entire human interactome showed the best performance showing the Area Under the Curve (AUC) of 0.93 for both ROC and precision-recall curves. A cutoff score of 0.7 was established as the high-confidence DMI prediction, resulting in a sensitivity of 66.3% and a specificity of 97.2% (**Figure 3.1 B**).

The pipeline outputs the DMI matches along with their scores with higher scores indicating a greater likelihood of being correct (**Figure 3.1 A**).

3.1.2 The application of the tool on HuRI PPI dataset

The developed DMI tool was applied to the HuRI dataset to detect PPIs potentially mediated by the DMI interface. Due to the inherent degeneracy of motifs, a large number of DMI matches were found within HuRI PPIs. After applying the cutoff of predictions with high confidence DMI match score (0.7), 13,406 high-confidence putative DMI interfaces are identified across 3,195 PPIs. Among these interactions, 54% had their top-ranked matches from the ligand (LIG) classes, and almost 20% DMI matches from the modification (MOD) class (**Figure 3.1 C**).

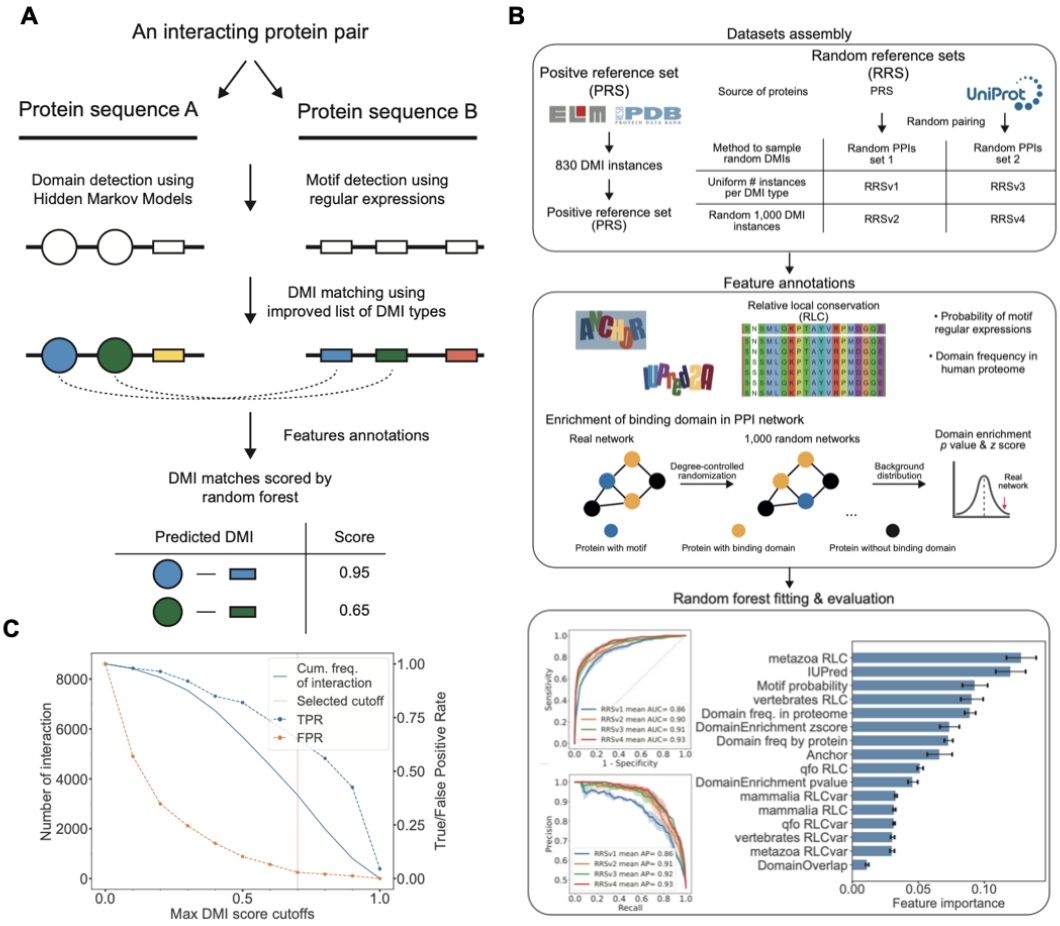


Figure 3.1: The development of DMI predictor and its application on HuRI. (A) Schematic illustrating the workflow of the developed DMI predictor. Here is the improved list of DMI types and trained Random Forest (RF) model incorporated into the DMI detection pipeline. (B) The top panel represents the assembly of PRS and different versions of RRS. The middle panel illustrates the annotation of features on the PRS and RRSs. The bottom panel represents the ROC curve of RF models trained using different sets of RRS. For each RRS version, ROC and PRS curves averaged across the triplicates of the RRS version were plotted by interpolation. ROC for the PRS curve of the RF models. The importance of different features to the RF trained using the PRS combined with the RRSv4 as quantified using mean decrease in impurity. (C) The developed DMI predictor was applied on PPIs that are detected in HuRI, and the scores of the predicted DMIs are titrated over increasing cutoffs. The dashed lines refer to the right y-axis, while the filled line refers to the left y-axis. The red vertical line implies the cutoff of 0.7 applied on the DMI scores to call a predicted DMI of high-confidence.

3.2 Integrating ClinVar mutation data with putative DMIs mapped on HuRI

The largest mutation database ClinVar (**see Chapter 1, section 1.1**) contains a comprehensive set of patient mutation data. My colleague processed this dataset by mapping mutations to proteins and applying

several filtering steps to the most recent version of ClinVar. The filtering process included only germline, non-synonymous single nucleotide variants (SNVs) with definitive clinical significance, excluding other variant types such as termination mutations. As a result, we have a total of 996,697 variants. Out of them, 45,035 are pathogenic, 73,806 are benign and 824,374 are variants of unknown significance (VUS).

The filtered variants were then overlapped with high-confidence domain-motif interfaces (DMIs) mapped on PPIs, focusing on those where at least one pathogenic or VUS variant falls within a predicted DMI. The PPI subset was visualized using a network tool, Cytoscape. We identified a total of 6,057 potential high-scored DMIs with at least one pathogenic or VUS mutation falling in the interface (**Figure 3.2 A**). As the subset is big for visualization and does not represent the details I zoomed out PPIs of HDAC4 and SPOP to show how it looks. Here HDAC4 has 6 partners with 6 high-confidence DMI predictions, where 5 partners might mediate the interaction through the LIG motif type interface and one interaction potentially occurs through the DOC type motif interface. Another protein SPOP has 6 interactions with 3 DEG and 3 DOC motif type interfaces (**Figure 3.2 A**). Among the DMIs in this subset, the most common SLiM type is LIG, with 2,867 instances, followed by MOD with 1,838 instances, and DOC with 881 instances. The least frequent SLiM types are TRG (304 instances), DEG (137 instances), and CLV (30 instances) (**Figure 3.2 B**).

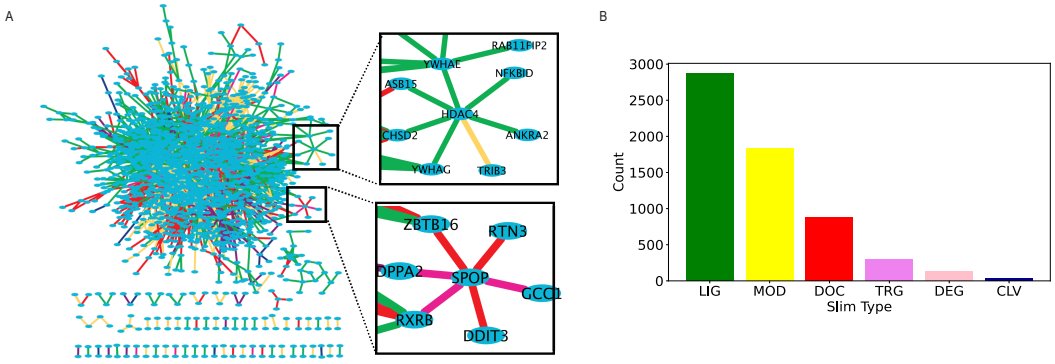


Figure 3.2: PPI network with predicted DMIs overlapped with ClinVar mutations. (A) PPI network illustrating the mapped predicted high-confidence DMIs with at least one pathogenic or VUS mutation overlapped. The blue nodes represent proteins, and the edges indicate the predicted DMI. The colors represent different SLiM types. (B) The bar plot illustrates the distribution of different SLiM types across the PPI network illustrated in A. Each SLiM type .

3.3 The data-driven approach to select disease-associated proteins and PPIs suitable for the experimental validation of DMIs

To select PPIs suitable for the experimental validation of putative DMIs I employed a data-driven approach annotating the PPIs with the subset with experimental features with information regarding available ORF sequences for these genes, which is important for candidate selection for experimental work.

For this, I explored our ORFeome collection database to gather information on the presence of the clone in the ORFeome collection. As it is essential to design an experiment close to native biological conditions, I selected full-length ORFs. Furthermore, the established pipeline implies the use of clonal ORFs to have a high success rate in cloning and sequence validation. Additionally, I assessed the number and types of mutations present at each interface mapped on PPIs.

Understanding the biological processes regulated by proteins encoded by these genes is also crucial. To do this I imported this information from UniProt and annotated the PPIs in the subset. Analyzing the candidates, I also checked how many partners these genes have. Given this biological information, I manually assessed the validity of the DMI prediction results. Some DMIs, despite having high match scores, did not align with current biological understanding. For example, we predicted an interface (DMI match score 0.741) involving WW domains and the DOC_WW_Pin1_4 motif between WWOX and MYOZ2. Pin1 is a multidomain protein with both a WW domain and a PPIase domain that work together to target specific sequences. The WW domain of Pin1 recognizes phosphorylated S/T-P motifs, while its PPIase activity regulates various cellular processes.

However, this prediction might be inaccurate because, although both Pin1 and WWOX contain WW domains and are involved in disease processes, their functions are distinct. Pin1's role as a PPIase with specific substrate targeting and isomerization activity sets it apart from WWOX, which does not perform isomerization but rather functions through protein interactions. The possibility that a highly-scored DMI might still be incorrect highlights the need for further refinement of the tool and underscores the importance of experimental validation, which is the next step in our proposed strategy.

As a result, I selected 31 annotated gene candidates. I applied the same approach and selected 105 gene partners. The selected candidates and partners form the network of 117 protein-protein interactions illus-

trated in (**Figure 3.3 A**). In this PPI network, 86 PPIs are mapped with predicted 88 domain-motif interfaces and 27 PPIs (found in HuRI) mediated by known 31 DMIs previously studied and annotated in the ELM database (**see chapter 1 section 1.2**) serve as positive controls for DMI validation. Since some candidates only had predicted interfaces, I included 4 additional partners where interactions (not found in HuRI) are mediated by known DMIs reported in the literature. Additionally, I included 78 partners that interact with the candidates via different interfaces, which will serve as negative controls.

3.3.1 Retestement of PPIs using BRET assay

We first cloned and sequence-verified the selected candidates to confirm protein expression after transfection into mammalian cells. If successful, we can then clone their interacting partners, making the cloning step more efficient. Our prior experience with the BRET assay indicated that proteins are better expressed when fused to Nanoluc (NL) luciferase at the N-terminus. Therefore, the candidates were genetically fused to the NL tag. Using the established cloning pipeline from Aim 1 (**see Chapter 1 section 1.5**) I successfully cloned ORFs for 19 candidate proteins. For the failed candidates, a second round of cloning was attempted. However, 3 ORFs yielded no growth in inoculation cultures, while sequencing of the remaining 8 showed either empty vectors or incorrect ORFs, suggesting that it may have happened due to cross-contamination. These results also showcase that the cloning step, particularly the manual picking of the colonies might lead to false-positive results.

For these successfully cloned ORFs, 114 partner ORFs were fused N-terminally to mCitrine, and 96 ORFs were successfully cloned, resulting in 96 PPIs available for detection in the BRET assay. As a result, we obtained significant BRET signals for 46 of these 96 PPIs with the valid expression of proteins (**Figure 3.3 B**). This retest rate surpasses those of gold-standard PPI datasets used in previous benchmarks of various binary PPI assays, including the BRET assay, highlighting the overall detectability of PPIs from HuRI (**Trepte et al. 2018**; **Braun et al. 2009**; **Choi et al. 2019**). We obtained significant BRET signals for 46 (48%) of these 96 PPIs with proteins expressed higher than the cutoff. This retest rate is notably higher compared to the retest rates of gold standard PPI datasets used in past benchmarking of various binary PPI assays, including this BRET assay, highlighting the enhanced detectability of PPIs from HuRI (**Trepte et al. 2018**; **Braun et al. 2009**; **Choi et al. 2019**).

Among these 46 PPIs, we selected 23 interactions involving 6 candidates (CTBP1, WWOX, PPP3CA, REPS1, SPOP, and IQCB1) for validating the predicted interfaces (**Figure 3.3 B**). The remaining 24 PPIs were not selected for further analysis because they involved 8 candidates with incomplete data, either missing known DMIs or consisting only of negative controls. For instance, PUF60 was detected with PPIs mediated solely by known DMIs or through different interfaces

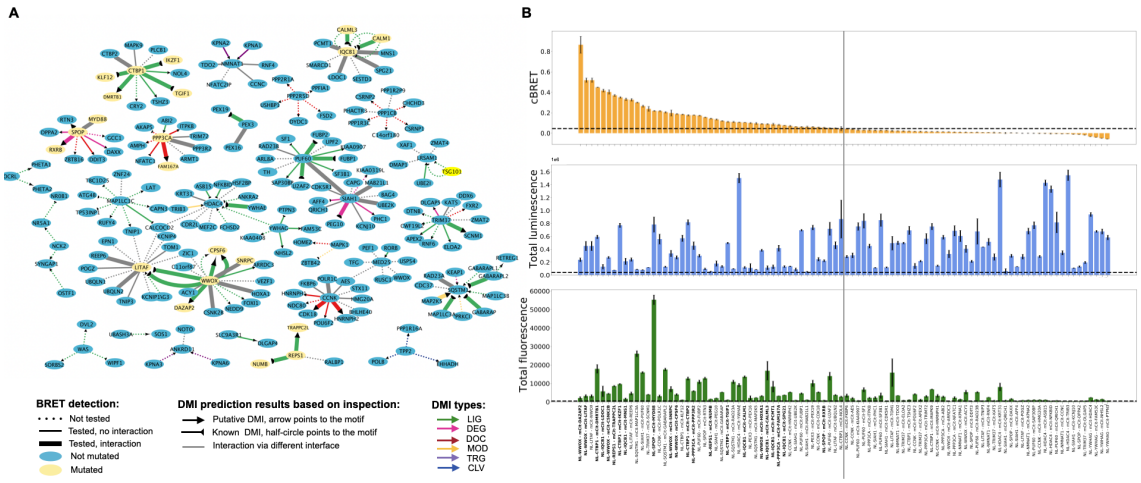


Figure 3.3: Experimental validation of predicted DMIs on PPIs. (A) PPI network illustrating selected DMI predictions and experimental retesting in BRET assay. (B) cBRET, total luminescence and fluorescence for 96 PPIs, where 31 PPIs have putative DMIs. Luminescence and fluorescence measurements indicate NL and mCit fusion protein expression levels, respectively. Black horizontal lines indicate expression level and PPI detection cutoffs. The gray vertical line separates the detected (left) from undetected PPIs. Protein pairs in bold indicate those selected for interface validation via site-directed mutagenesis. Error bars indicate STD of three technical replicates.

To design mutants for the experimental DMI validation and variants from patients (**see section 3.4)** I used the predicted interface AF-MM structures run by my colleague. I visualized the predicted structures with the protein structure visualizing tool, PyMol to guide the design. We manually designed single point mutations at potential motif and domain sites of interacting protein pairs, along with deletions of motifs or regions, resulting in 2-4 mutations per motif and domain. In total, we designed 55 mutations that fall into the predicted DMI and likely disrupt the interaction.

Next, I cloned the designed mutations using adapted to medium-throughput site-directed mutagenesis (**see Appendix 5.1.2**) and successfully cloned 44 mutants, 18 for domain and 27 mutants for motif validation. The expression of the mutated proteins was tested and compared to wild-type proteins (**see Appendix, Figure 5.1**). Mutants with low expression (e.g. motif deletion of LITAF) might interfere less

or not at all with the protein, potentially leading to false negative results. Consequently, these low-expressing mutants were excluded from further validation.

The successfully cloned and expressed protein candidates and their partners were used further for DMI validation using BRET saturation assay (**Trepte et al. 2018**; **Lee et al. 2024**). In this assay, I generated mutated constructs and performed a donor saturation experiment, where the amount of NL-candidate ORF construct (1 and 2ng) encoding NL-fused proteins, were co-transfected with increasing amounts of mCit-partner ORF (12.5, 25, 50, 100, 200 ng) encoding mCitrine-fused proteins performing in total 6 measuring points. Thus, with an increased concentration of acceptor protein, the BRET signal should increase until it attains a saturation value called maximum BRET. This saturated BRET value is reached when all the donor molecules interact with the acceptor molecule.

3.3.2 Testing the localization of the wild-type proteins and mutants using Bioluminescence Imaging

As was mentioned in section 1.5 the disruption of protein-protein interaction may happen due to the mislocalization of the mutant rather than the effect of the mutant on the interaction. One of the advantages of the BRET assay is that the tags for interaction testing can also be used to monitor protein location within the cell and the BRET signal can even be visualized in live cells via bioluminescence imaging, shortly BLI (**Goyet et al. 2016**; **Kobayashi et al. 2019**). It was also shown that it can be scaled up using a high-content screening (HCS) microscopy (**J. Kim et al. 2016**). Thus, with the support of the microscopy core facility at IMB, we were motivated to perform BLI by using a 96-well plate format on an HCS microscope, named Opera Phenix.

To do this, we selected some of those mutants for DMI validation (TGIF1_24_28del, DMRTB1_21_25del, CPSF6_323_327del, FAM167A_3_9del) as well as patient variants (DMRTB1_R25H, WWOX_H37D,LITAF_Y61D, FAM167A_V8M) that showed the effect on the binding affinity of the interactions compared to the wild-type **(see subsection 3.3.3)**. The selected mutants and variants, paired with wild-type partners at a ratio of 10:10 ng, were transfected into pre-seeded U2OS cells in a 96-well plate using Fugene as the transfection agent. Upon transfection, cells were incubated for 24 hours. The following day, DRAQ5 and CellMask dyes were applied to stain the nucleus and cytoplasm, respectively (data not shown), and the cells were imaged immediately using the Opera Phenix system. Initially, fluorescence

was imaged in each well. To detect luminescence, furimazine substrate (from the Nano-Glo kit) was added to the wells, enabling the oxidation of NanoLuc luciferase for luminescence detection.

Below, I will first discuss the results of validating predicted interfaces and microscopy data. For the negative controls, which lack resolved structures, we employed the AF-MM fragmentation approach (**see Chapter 2, Article II**) to predict potential interfaces. This method helps us infer interaction sites in the absence of structural data, providing insights into the validity of our predictions and the reliability of the negative controls.

3.3.3 Validation of DMI predictions

Experimental validation of interfaces involving CTBP1 interactions

CTBP1 is a transcriptional co-repressor. Unlike many transcription factors, CTBP1 does not directly bind DNA (**Filograna et al. 2024**; **Valente et al. 2013**). Instead, it interacts with transcription factors through a hydrophobic cleft in its substrate-binding domain, which recognizes the PxDLS motif. This cleft is crucial for recruiting other corepressor components such as histone deacetylases (HDACs), methyltransferases (HMTases), and additional transcriptional repressors necessary for its repressor activity (**Filograna et al. 2024**; **Valente et al. 2013**). For the CTBP1 candidate, we have cloned the partners with the same interface LIG_CtBP_PxDLS_1 class, TGIF1 and IKZF1 with known DMIs, partner DMRTB1 with predicted interface and CTBP2 as a negative control, meaning that this interaction likely happens through a domain-domain interface.

CTBP1-TGIF1

CTBP1 binds to the PLDLS motif of the transcription factor, TGIF1 (**Figure 3.4 A**). This interface has been functionally studied and annotated in ELM (**Melhuish 2000**), but no crystallized structure is available. The predicted AF-MM structure with a high confidence score of 0.8, suggests that the proline at position 24 of TGIF1 fits well into the hydrophobic pocket of CTBP1. Furthermore, two leucines contribute to beta augmentation, allowing the sidechain of the motif to enter a deep hydrophobic groove. In addition, a negatively charged aspartate is in proximity to phenylalanine, a non-polar hydrophobic residue (**Figure**

3.4 B). This suggests that phenylalanine's aromatic ring might be involved in pi-stacking interaction by stabilizing the interface.

I mutated residues A41 and C27 in the CTBP1 binding pocket that interacts directly with the motif, as well as a residue K54A, which is away from the motif and likely will not affect the binding (**Figure 3.4 B**). Additionally, we deleted the motif from TGIF1 to potentially completely disrupt the interaction (**Figure 3.4 A**). The BRET data showed that mutations A41D, C27E, and C27D in CTBP1 completely disrupted the interaction with TGIF1 (**Figure 3.4 E**), whereas the K54A mutation did not disrupt the binding. The deletion of the motif in TGIF1 showed the loss of interaction. The expression data is shown in the Appendix (**Figure 5.2 A**). The microscopy data suggests that the mutant with the removed motif is localized similarly to the wild-type (**Figure 3.5**)

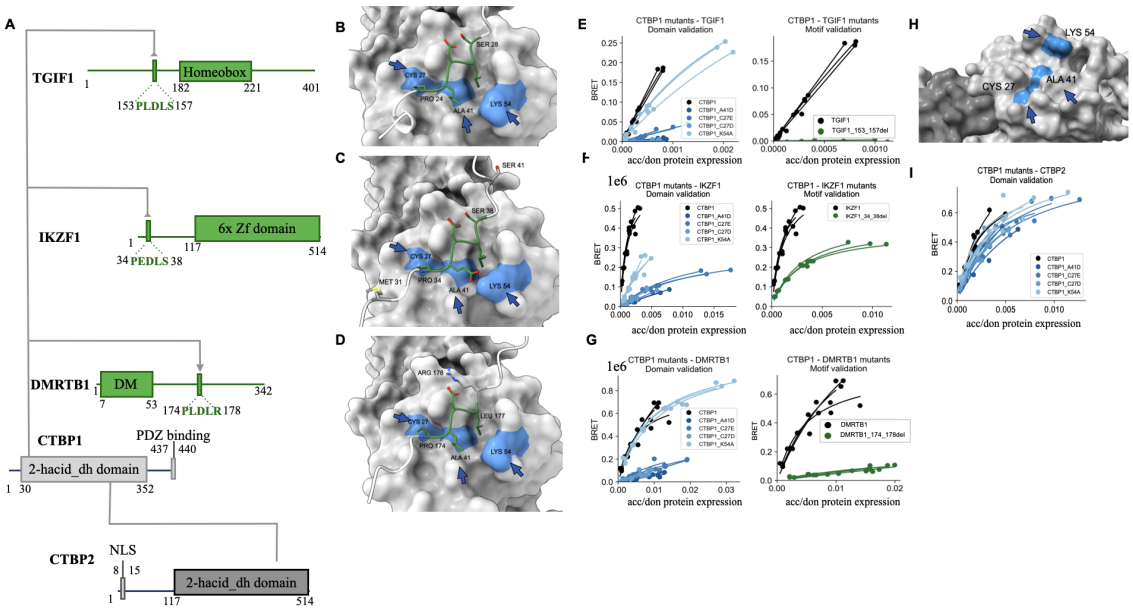


Figure 3.4: DMIs mediating CTBP1-centric PPIs (A) The schematic illustration of CTBP1 interactions mediated by predicted LIG motif type (green) predicted (the arrow end pointing to motif) and known (half-circle end pointing to motif) DMIs, and negative control (gray) interaction mediated by different interface, DDI. (B-D)Predicted by AF-MM interface interaction structures of CTBP1 with TGIF1 (B, known DMI), IKZF1 (C, known DMI), and DMRTB1 (D, predicted DMI). The interacting CTBP1 domain (gray) with highlighted residues (blue), mutated for domain validation, is shown. Motifs (green) and flanking regions (white) are indicated for each interaction. (E-G) Experimental confirmation of known DMIs (E, CTBP1-TGIF1), (F, CTBP1-IKZF1) and validation of putative DMI (G, CTBP1-DMRTB1) using saturation assays, with BRET measured as a function of acceptor/donor expression ratio. The left panels show saturation curves for wild-type CTBP1 and single-point mutants (A41D, C27E, C27D, K54A) for domain validation. The right panels display binding curves for the wild-type partner proteins (TGIF1, IKZF1, and DMRTB1) and their mutants with deleted motifs. (H-I) Predicted structure of the negative control PPI using the AF-MM fragmentation approach (H), where CTBP1 (gray) and CTBP2 (dark gray) interacting domains are shown, with CTBP1 residues (blue) mutated for domain validation and experimental validation of the CTBP1 domain being part of the DDI interface between CTBP1 and CTBP2 (I).

CTBP1-IKZF1

Another notable interaction involves CTBP1 and the PEDLS motif of the transcription factor IKZF1 (**Figure 3.4 A**). It is a DNA-binding protein that regulates transcription through association with HDAC-dependent and independent complexes. The previous study tested if the conserved PEDLS motif in IKZF1 was crucial for this interaction by creating mutations that either deleted this sequence or altered its core amino acids (**Koipally 2000**). The mutated IKZF1 proteins failed to

bind CTBP1, confirming the importance of the PEDLS motif for their interaction. Similar to TGIF1, the motif was predicted to fit the cleft of the CTBP1 domain (**Figure 3.4 C**). However, negatively charged glutamate on the IKZF motif and positively charged lysine on CTBP1 forming an electrostatic interaction. Therefore we expect that K54A might slightly affect the binding.

Validation with the same CTBP1 mutants and deletion of the IKZF1 motif showed only partial disruption of the interaction. This partial perturbation suggests that while the PEDLS motif is crucial, other factors may also contribute to the binding stability (**Figure 3.4 F**). For example, IKZF1 contains zinc finger domains essential for DNA binding and dimerization (**Figure 3.4 A**), which might still interact with CTBP1 through these domains. This hypothesis needs to be tested with additional downstream experiments such as mutation of zinc finger domains of IKZF1 protein. The obtained microscopy data indicate that the localization of the mutant with the removed motif in IKZF1 is similar to the wild-type protein (**Figure 3.5 B**).

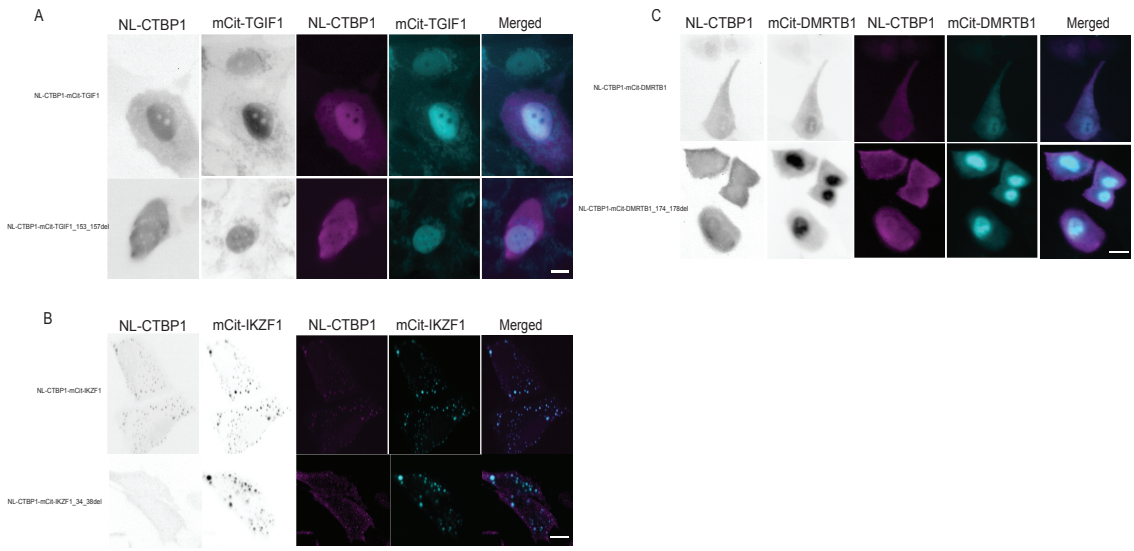


Figure 3.5: The localization of wild-type and mutants. Bright-field microscopy image of U2OS cells showing luminescence (magenta) indicating the presence of NL-CTBP1 and fluorescence intensity (cyan) of mCit-TGIF1. The images depict the localization of the wild-type proteins (top panel) and the mutant with the removed motif (bottom panel) relative to the wild-type. Scale bar = 10 µm.

CTBP1-DMRTB1

The same domain of CTBP1 was predicted to bind a PLDLR motif of DMRTB1 with a high score of 0.881. This motif is annotated in ELM

but is found in murine HDAC9. However, this potential interface between CTBP1 and DMRTB1 has not been discovered yet (**Figure 3.4 A**). The AF-MM structure also looks very promising. The PLDL part of the motif fits the hydrophobic pocket of CTBP1 similar to known SLiM instances mentioned earlier, while positively charged arginine residue and glutamate on the domain form salt bridges that contribute to the stabilization of the interaction (**Figure 3.4 D**). The BRET data supports the prediction findings, with domain mutations significantly reducing the binding and the deletion of the motif leading to the loss of interaction (**Figure 3.4 G**). We also showed that DMRTB1 mutant is localized similarly to the wild-type protein (**Figure 3.5 C**), while the expression data shows that the Depression of DMRTB1 is slightly higher compared to the wild-type (**see Appendix, Figure 5.2 A**).

CTBP1-CTBP2

As a negative control, we confirmed the PPI of CTBP1 with CTBP2. Although CTBP1 and CTBP2 proteins share 78% amino acid identity and 83% similarity, there are slight differences in their sequences that contribute to their distinct functions (**Ding et al. 2020**). For example, CTBP2 has a nuclear localization signal at the N-terminus but lacks a PDZ-binding domain. Previously, it was demonstrated that both CTBP1 and CTBP2 contain an NADH-dependent homo- and heterodimerization domain, which facilitates dimerization in response to increased NADH levels (**Figure 3.4 A**). This dimerization further promotes the nuclear retention of CTBP1.

Currently, there is no resolved structure available for the interaction interface of CTBP1 and CTBP2. To address this, we employed a fragmentation approach using AF-MM. The AF-MM prediction was consistent with previous studies, suggesting that the domains of CTBP1 interact with CTBP2 to form a dimer (**Figure 3.4 H**). BRET assay indicated that single-point mutants on the PxDLS-binding cleft do not disturb the binding, suggesting that this cleft is not essential for the dimerization of CTBP1 and CTBP2, which is also in line with the predicted structural model (**Figure 3.4 I**).

Experimental validation of interfaces involving WWOX interactions

WWOX is a putative oxidoreductase: it has two WW domains (WW-1 and WW-2) maintaining many interactions, NLS and an SDR (steroid

dehydrogenase) domain involved in metabolism. DMI predictor tool predicted that WWOX binds via the same interface LIG_WW_1 and LIG_WW_3 SLiM classes with LITAF (known as two DMI interfaces), CPSF6 (three interfaces), and DAZAP2 (one DMI). Additionally, negative control partners HOXA1, CSNK2B and SNRPC are used.

WWOX-LITAF

LITAF plays a role in endosomal protein trafficking and targets proteins for lysosomal degradation (**Lee et al. 2011**). It consists of two short PPSY motifs at the N-terminus and SLD domain with a hydrophobic cysteine-rich core region anchored to the membrane of the lysosome (**Figure 3.6 A**). Previously, it found that the WW-1 domain binds specifically to PPSY motifs in LITAF, whereas the WW-2 domain does not (**Ludes-Meyers et al. 2004**).

Using AF-MM, we predicted a high-confidence (0.8) structural model of the interface between the first motif (20-23) and tandem WW domains. The structure suggests this motif is recognized by WW-1 (**Figure 3.6 B (i)**). We also predicted the structure (with the same confidence score) of the second known interface of the second PPSY motif (58-61) and tandem WW domains. Similarly, the second motif prefers interaction with the WW-2 domain (**Figure 3.6 B (ii)**). The prolines and tyrosine residues on the motif fit into the pocket WW1 containing tryptophan and tyrosine (**Figure 3.6 B**).

The prediction indicates that both motifs might interact with the WW1 domain, though they bind in the same manner, suggesting multivalency, where multiple interactions between identical (by sequence) motifs and one domain occur. To confirm these interfaces, we designed mutated residues on the WW1 domain and motif (**Figure 3.6 B**). In addition, I also generated motif deletions, each separate and N-terminal part. However, the expression levels were lower than the threshold (**see Appendix Figure 5.2 B**) and we excluded these deletions for further study.

Experimental validation showed partial disruption of binding with mutations Y33H, Y33D, and W44K in WW1, while mutations on prolines and tyrosines in the motifs had varying effects. Replacement of tyrosines in motifs with aspartate completely disrupted binding (**Figure 3.6 F**).

In contrast, Ludes-Meyers et al. (2004) demonstrated that mutating tyrosine to alanine on the first motif significantly reduces binding, while mutating tyrosine to alanine on the second motif does not affect binding. Alanine is a small, non-polar amino acid that lacks the aromatic

side chain of tyrosine. The loss of this aromatic interaction might significantly reduce but not eliminate the binding, as observed in Meyers's study.

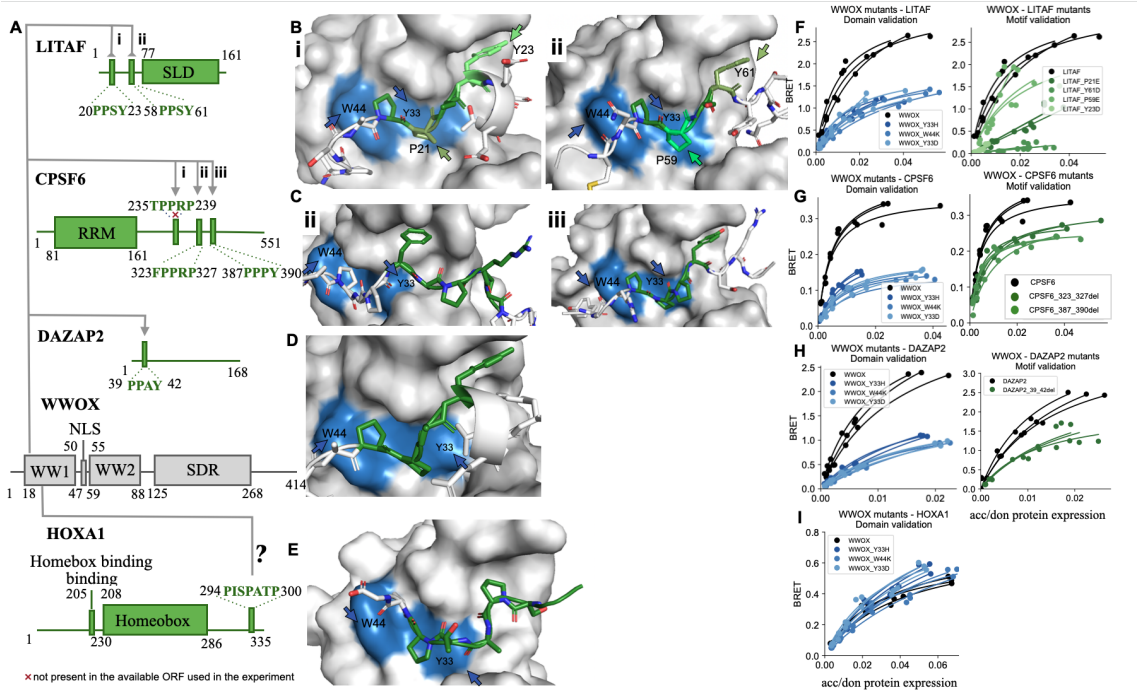


Figure 3.6: DMIs mediating WWOX-centric PPIs (A) Schematic illustration of PPIs mediated by DMIs. The edge ending points towards the predicted motif, where the arrow implies predicted DMI, while the half-circle points to the known DMI and gray indicates interaction mediated by different interfaces, where the question means that this interface was predicted by AF-MM using fragmentation approach. (Bi) Predicted interface interaction structure of the WW1 domain with the first PPSY motif in the WWOX-LITAF interaction. The structure highlights mutated residues on the domain (in blue) and on the motif (in green), with arrows pointing to these residues. (Bii) Predicted interface interaction structure of the WW-1 domain with the second PPSY motif of WWOX-LITAF interaction. (Cii) Predicted structure illustrating the second motif on CPSF6 and tandem WW domains as shown in A scheme. (Ciii) Predicted interface interaction structure of the WW-1 domain with the third motif of WWOX-CPSF6 interaction. (D) The putative model of the motif on DAZAP2 and tandem WW domains. (E) Predicted interface of the negative control PPI. (F) Experimental confirmation of known DMIs of WWOX-LITAF using BRET saturation assay. (G) Experimental validation of predicted DMIs of WWOX-CPSF6 using BRET saturation assay. (H) Experimental validation of putative DMI of WWOX-DAZAP2 using BRET saturation assay. (I) Experimental validation of whether the domain is involved or not in the interface of the negative control.

WWOX-CPSF6

DMI predictions identified three potential interfaces between WWOX and CPSF6: one with the PPPY motif and two with the TPPRP and FPPRP motifs located at the C-terminus of CPSF6 (**3.6 A**). One study

identified several novel interactions involving WW domains using mass spectrometry. They found that CPSF6 is associated with the WW-1 domain of WWOX. They further investigated whether specific proline-based peptide motifs are present in proteins bound by WW domains and found that CPSF6 contains PPPY motif as a potential interface between WWOX and CPSF6. However, no validation of this interface was done in this study (**Ingham et al. 2005**).

AF-MM predictions indicated that the FPPRP motif binds to the WW2 domain (**Figure 3.6 C (ii)**), while the PPPY motif is more between WW domains (**Figure 3.6 C (iii)**). BRET experiments showed that single-point mutants of residues on the WW-1 domain significantly disrupted the binding. (**Figure 3.6 G, right panel**). Moreover, the mutant with removed motif FPPRP on CPSF6 partially disrupted the binding, similar to the effect of the mutant with the deleted third motif, PPPY, on CPSF6 (**Figure 3.6 G, left panel**). Given our predictions and experimental data, one might speculate that the mutant with all removed motifs might completely disrupt the binding.

WWOX-DAZAP2

The same DMI interface of LIG_WW_1 class was predicted by the DMI predictor tool with a DMI match score of 0.9 for tandem WW domains of WWOX and PPAY N-terminal motif in DAZAP2 (**Figure 3.6 A**). AF-MM model of the interface proposes the PPAY motif to fit well the hydrophobic groove formed on the WW-1 domain. In the WW1 domain, the tryptophan residue (W44) and tyrosine residue (Y33) are positioned in a way that allows aromatic stacking with the proline residues of the PPAY motif (**Figure 3.6 D**). The involvement of the WW-1 domain in the predicted interface was experimentally validated, demonstrating the reduction in binding of domain mutants (**Figure 3.6 H, left panel**). The deletion of the predicted motif on DAZAP2 slightly affected the interaction with WWOX (**Figure 3.6 H, right panel**).

WWOX-HOXA1

I used HOXA1 as the negative control, assuming the interaction is mediated via a different interface (**Figure 3.6 A**). No DMI prediction was found on this interaction upon the application of the DMI predictor tool. HOX1 does not contain PPxY (where x represents any amino acid), PPLP or xPPRX motif recognized by WW domains. Using the fragmentation approach, my colleague predicted the potential interface. The WW tandem domain in WWOX is modeled with the disordered re-

gion 294-302 of HOXA1 with moderate confidence (pLDDT 73) (**Figure 3.6 E**). Here the 294-300 (PISPATP) of HOXA1 matches the regex of LIG_SH3_3 that binds to the SH3 domain. According to the predicted structure, two prolines at the C terminal of the peptide stack nicely with aromatic sidechains (W and Y) from the WW domains in a similar way as the LIG_WW_1 class. However, BRET experiments showed WW1 mutants did not change the binding with HOXA1, meaning that this domain might not interact (**Figure 3.6 K**). This data also showed the limitation of AF-MM in specificity.

WWOX-SNRPC

Another negative control is WWOX and its partner SNRPC in BRET (see Appendix). However, we had a DMI prediction that slightly scored below the cutoff, where the LIG_WW_3 motif within the proline-rich region of SNRPC was predicted to bind to the WW1 domain. The titration studies showed that the mutant constructs on the domain and the deletion of the potential motif as well as the whole proline-rich region left the interaction intact. With these conditions, we could not validate this interface prediction (**see Appendix, Figure 5.3 C & D**). We also tried the fragmentation approach using AF-MM, where the only promising prediction that survives the cutoff is an ordered-ordered pair. The prediction involves the Zn finger from SNRPC and the C-terminal SDR domain from WWOX (**see Appendix, 5.3 A**), but we did not test this prediction experimentally. Given these findings, the DMI predictor returned interface prediction, suggesting that the GPPRP motif of LIG_WW_3 class binds to WW domains is likely wrong. This motif is recognized by group III WW domains, whereas WWOX contains WW domains from group I. Our structural data also showed that predicted.PPR. motifs from class LIG_WW_3 class are predicted to be positioned away from the binding groove. These data also point to the inability of DMI predictors to discriminate domain preference within domain class.

WWOX-CNSK2B

This PPI was also annotated as the negative control. Similar to HOXA1, CSNK2B does not have proline-rich stretches and we did not have any DMI predictions. AF-MM prediction was not done. BRET signals of the mutant did not differ from the wild-type protein titration results (**see Appendix 5.3 E & F**).

Experimental validation of interfaces involving IQCB1 interactions

IQCB1 contains a tyrosine phosphorylation site, a coiled-coil region, and three helical calmodulin-binding motifs. The calmodulin-binding motif is a ligand type motif with the consensus [I,L,V]QxxxRGxxx[R,K] with characteristic residues being a hydrophobic residue at position 1, highly conserved glutamine at position 2, basic charges at positions 6 and 11, and a variable glycine at position 7. Two of these motifs are known and also annotated in the ELM database (321-336, 391-407) (**X. Luo et al. 2005**). Whereas the third motif (298-314) was predicted by the DMI predictor (**Figure 3.7 A**). The DMI tool predicted these motifs interact with the EF-hand repeat domains of CALM1 and CALML3 proteins.

IQCB1-CALM1

The DMI predictor gave a high DMI match score of 0.9 and found these motifs potentially interacting with the tandem EF-hand domains of CALM1. Upon binding of four Ca ions through these motifs, CALM1 changes its conformation from a closed form to an open one, exposing a hydrophobic surface capable of interacting with different target proteins. AF-MM predicted the interface of the third motifs and the tandem EF-hand domains (**Figure 3.7 B**). The predicted model suggests that the IQCB1 motif is tightly wrapped and embedded within the binding pocket of CALM1. Validation experiments were limited due to the non-canonical isoform 2 of the IQCB1 clone, which lacks two full first and second motifs. We had one successful mutant E120H for domain validation which is predicted to be away from the IQCB1 motif (**Figure 3.7 B (iiia)**). However experimental data showed that E120H did not likely affect the binding with IQCB1. In contrast, the deletion of the motif partially reduced the interaction (**Figure 3.7 E**), suggesting that while the motif is important, other factors or regions may also play a role in maintaining the overall interaction between proteins.

IQCB1-CALML3

Similar to the DMI mediating interaction IQCB1-CALM1, it was predicted that EF-hand domain-containing CALML3 likely recognizes the same motifs of IQCB1 (**Figure 3.7 A**). The AF-MM prediction showed a similar outcome (**Figure 3.7 C**). The BRET data supports the prediction showing that E85K and the deletion of the motif weakened the binding with wild-type protein pair (**Figure 3.7 F**).

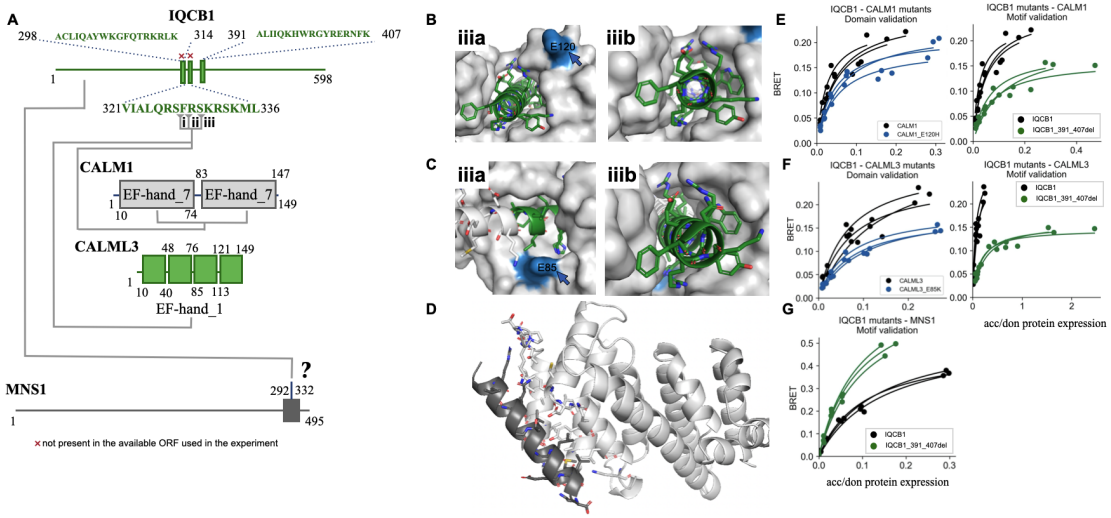


Figure 3.7: DMIs mediating IQCB1-centric PPIs (A) Schematic illustration of PPIs mediated by DMIs. The edge ending points towards the predicted motif, where the arrow implies predicted DMI, while the half-circle points to the known DMI and gray indicates interaction mediated by different interfaces, where the question means that this interface was predicted by AF-MM using the fragmentation approach. (Biii a&b) Predicted interface interaction structure of the known DMI, tandem Eh domain in contact with the third motif in the IQCB1-CALM1 interaction. The structure highlights mutated residues on the domain (in blue) and on the motif (in green), with arrows pointing to these residues. Ciii a&b) Predicted interface interaction structure of the known DMI, tandem Eh domain in contact with the third motif in the IQCB1-CALML3 interaction. The structure highlights mutated residues on the domain (in blue) and on the motif (in green), with arrows pointing to these residues. (D) Predicted novel interface of the negative control PPI using AF-MM fragmentation approach. (E) Experimental confirmation of known DMIs of CTBP1-CALM1 using BRET saturation assay. (F) Experimental confirmation of known DMIs of CTBP1-CALML3 using BRET saturation assay. (G) Experimental testing of the third motif being involved or not in the interface of the negative control.

IQCB1-MNS1

One of the negative controls is the interaction of IQCB1 with MNS1. It has no folded globular region, the monomeric structure of it shows the protein is composed of long helices. There was no putative DMI returned using the DMI predictor tool. To test and verify that the motif is not part of the interface of the interaction with MNS1, I tested the motif with the removed motif in pair with the wild-type MNS1. Interestingly, BRET data showed that the deletion of the motif caused an increase in BRET.

Using the AF-MM fragmentation approach, the predicted model suggests that helices of IQCB1 potentially bind to the C-terminal disordered region of MNS1, 292-332 (**Figure 3.7 A and D**). Despite a very high predictive score (0.89) manual inspection of the predicted interfaces of

fragments from the same region shows AF putting the fragments at different sites (not shown). Therefore, this putative interface might be wrong. The expression data is shown in **Appendix 5.4**

Experimental validation of interfaces involving PPP3CA interactions

PPP3CA is the phosphatase of type PP3, (its old name is calcineurin or PP2B) that recognizes its substrates via DOC_PP2B motifs. There are 3 catalytic subunits (PPP3CA, PPP3CB, PPP3CC) and two regulatory subunits (PPP3R1, PPP3R2). Upon increase in Ca2+ levels, it forms a complex composed of calcineurin A (catalytic subunit that is dependent on calmodulin) and a regulatory Ca2+-binding subunit (calcineurin B).

PPP3CA-FAM167A

The motif of DOC_PP2B_PxIxI_1 class in FAM167A (3-9 aa) is perfectly predicted by our DMI predictor tool to bind to the calcineurin (Metallophos) domain in PPP3CA (**Figure 3.8 A**) AF-MM putative structure predicts the potential motif forms the contacts along the edge of two beta sheets in the calcineurin PPP3CA (**Figure 3.8 C**). The results showed that mutants on the domain reduced the binding , while the deletion of the motif of FAM167A completely disrupted the interaction, while the expression of the wild-type and mutants were above the cutoff (**Figure 3.8 E**). Taken together, it can be suggested that FAM167A might be a potential substrate for PPP3CA.

PPP3CA-PPP3R2

PPP3CA interaction with PPP3R2 is mediated by different interfaces. PPP3R2 is the regulatory subunit that binds calcium ions and modulates the activity of PPP3CA in response to changes in intracellular calcium levels. PPP3R2 contains EF-hand domains. When intracellular calcium concentrations rise, calcium binds to these domain repeats in PPP3R2, inducing conformational changes that activate PPP3CA (**Figure 3.8 D**).

This PPI serves as a negative control in this study (**Figure 3.8 A**). BRET signals for the single mutants on the PPP3CA domain did not affect the interaction, potentially meaning that these residues of this domain might not contact PPP3R2 (**Figure 3.8 F**). Microscopy

data suggests that the deletion of the motif did not change localization compared to the wild-type (**Figure 3.8 B**).

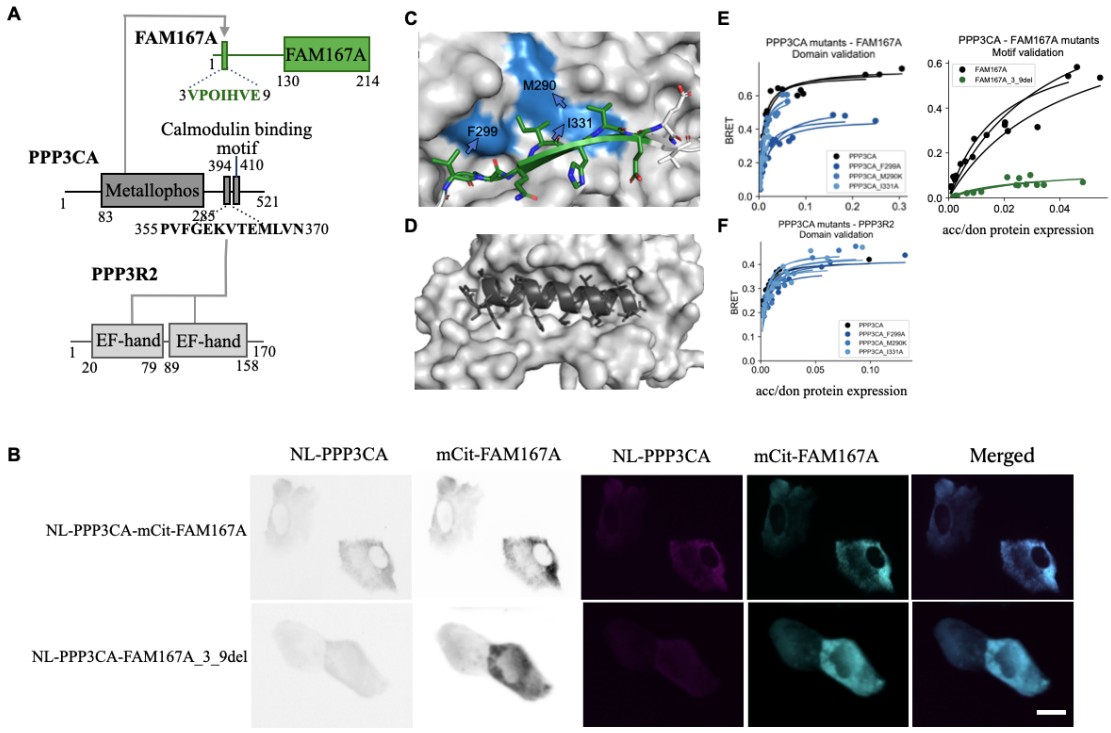


Figure 3.8: DMIs mediating PPP3CA-centric PPIs (A) Schematic illustration of PPIs mediated by DMIs. The edge ending points towards the predicted motif, where the arrow implies predicted DMI, while the half-circle points to the known DMI and gray indicates interaction mediated by different interfaces, where the question means that this interface was predicted by AF-MM using fragmentation approach. (B) The localization of wild-type and mutants. Bright-field microscopy image of U2OS cells showing luminescence (magenta) indicating the presence of NL-PPP3CA and fluorescence intensity (cyan) of mCit-FAM167A. The images depict the localization of the wild-type proteins (top panel) and the mutant with the removed motif (bottom panel) relative to the wild-type. Scale bar = 10 µm. (C) Predicted interface interaction structure of the predicted DMI, tandem Metallophos domain in contact with the motif in the PPP3CA-FAM167A interaction. The structure highlights mutated residues on the domain (in blue) and on the motif (in green), with arrows pointing to these residues. (D) Predicted known interface of the negative control PPI using AF-MM fragmentation approach. (E) Experimental validation of putative DMIs of PPP3CA-FAM167A using BRET saturation assay. (F) Experimental testing of the domain is involved or not in the interface of the negative control.

Experimental validation of interfaces involving SPOP interactions

SPOP is the component of RING-based BCR (BTB-CUL3-RBX1) E3 ubiquitin-protein ligase complex that mediates ubiquitination of targeted proteins, leading to proteasomal degradation. It contains two

globular domains MATH and BTB domains. Cullin E3 ligase binds to the BTB domain while the MATH domain directly recruits the substrates of the E3 ligase complex for ubiquitination. In complex with Cul3, the binding of SPOP to the motif leads to the proteasomal degradation of the substrate.

SPOP-RXRB

The DMI tool predicted the MATH domain might bind to two motifs of the DEG_SPOP_SBC_1 class at the N-terminal region of RXRB protein with a DMI match score of 0.899. RXRB also contains four Zn finger repeats and an LBD domain (**Figure 3.9 A**). There is no solved structure of this interaction interface is resolved. The AF-MM model suggests that the SPOP and RXRB interface is promising, with the motif docking into a hydrophobic cleft on the SPOP domain (**Figure 3.9 C**).

BRET experiments involved testing four mutants in SPOP: G132Q and F102V core mutants significantly reduced binding, whereas S119L and R70T edge mutants did not affect the interaction (**Figure 3.9 H**). Interestingly, both the deletion of the first motif and the deletion of both motifs resulted in BRET signals similar to the wild-type interaction, indicating that the interaction remained intact (**Figure 3.9 H**). The obtained findings suggest that the predicted motifs of RXRB were not verified with the deletion of the motifs, and the prediction of this interface is likely to be wrong.

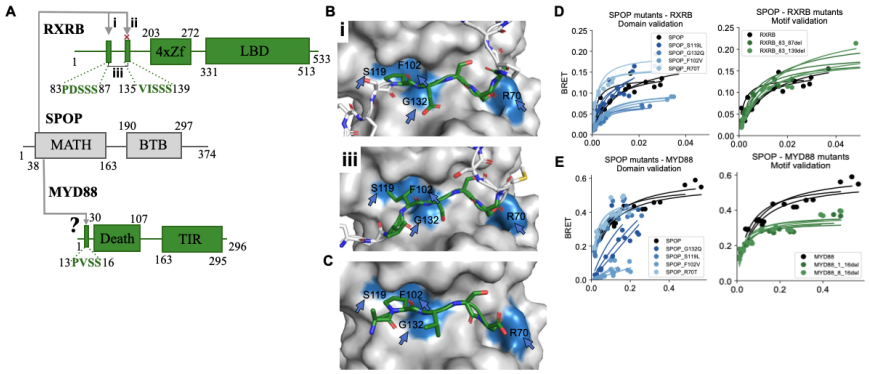


Figure 3.9: DMIs mediating SPOP PPIs (A) Schematic illustration of PPIs mediated by DMIs. The edge ending points towards the predicted motif, where the arrow implies predicted DMI, while the half-circle points to the known DMI and gray indicates interaction mediated by different interfaces, where the question means that this interface was predicted by AF-MM using fragmentation approach. (Bi) Predicted interface interaction structure of the predicted DMI, where the domain is in contact with the first motif in the SPOP-RXRB interaction. The structure highlights mutated residues on the domain (in blue) and on the motif (in green), with arrows pointing to these residues. (Biii) Predicted interface interaction structure of the predicted DMI, where the domain is in contact with the second motif in the SPOP-RXRB interaction. The structure highlights mutated residues on the domain (in blue) and on the motif (in green), with arrows pointing to these residues. (C) Predicted novel interface of the negative control PPI using AF-MM fragmentation approach. (D) Experimental validation of putative DMIs of SPOP-RXRB using BRET saturation assay. (E) Experimental testing of whether the domain and motif are involved or not in the interface of the negative control.

SPOP-MYD88

Another interaction partner of SPOP is MYD88. This partner has two globular domains Death and TIR. Slim DEG_SPOP_SBC_1, has been detected in region 12-19 (APVSSTSS) of MYD88, with DMIMatchScore 0.55. The DMIMatchScore is below the 0.7 cutoff that we set, therefore this PPI is treated as a negative control (**Figure 3.9 A**). Overlapping fragments that cover the core binding motif are also repeatedly modeled at the same interface with high confidence, making the interface very likely to be true. The core motif is likely 13-16 PVSS.

Taking the biological function of the proteins we hypothesized that this interface might be true. Therefore we tested the previously mentioned mutants for domain validation and the core mutants significantly perturbed the interaction. On the other side, we removed the motif and N-terminal part of MYD88 and obtained unexpected findings. The BRET experiments show that the deletion of the N-terminal part led to the lower BRET, but enhanced the binding affinity (**Figure 3.9 I**).

Based on our observations, we hypothesize that the deletion of the N-terminal part of MYD88 might changed the spatial rearrangement of the proteins, increasing the distance between the donor and acceptor fluorophores or altering their orientation. The increased distance between tags is indicated by a lower BRET signal. At the same time, this deletion might increase the accessibility of the binding site for SPOP leading to enhanced binding affinity. Later I found that the previous study reported that the co-IP and ubiquitination assay showed that MyD88-VSSTS mutant still binds to SPOP and can be ubiquitinated by SPOP at levels comparable with those of wild-type MyD88. Moreover, they reported that an SBClike motif (146-VDSSV-150 aa) located in the middle of MyD88 is indispensable for MyD88–SPOP interaction and SPOP-dependent ubiquitination (**Li et al. 2020**).

Experimental validation of interfaces involving REPS1 interactions

REPS1- NUMB

REPS1 contains tandem repeats of EH domains. EH domains are exclusively found in proteins that function in endocytosis and vesicular trafficking and are believed to regulate these processes. They recognize proteins containing single or multiple NPF (Asn-Pro-Phe) motifs, like NUMB (**Figure 3.10 A**). In ELM the canonical EH binding peptide is a strongly conserved NPF motif. NUMB also contains PID and NUMB domains at the N-terminal and middle part of the protein. This interface is known. Proline and Phenylalanine fit the hydrophobic pocket on the EH-domain very well according to the predicted AF-MM structure (**Figure 3.10 C**). Although BRET experiments demonstrated that W275A did not significantly affect binding (**Figure 3.10 E**), the expression of the L271D mutant was destabilised while being co-expressed with wild-type NUMB **Appendix, Figure 5.4 D**. Similarly, the deletion of the motif as single mutants was not expressed well in my hands. Therefore, it was hard to make any conclusions regarding the interface.

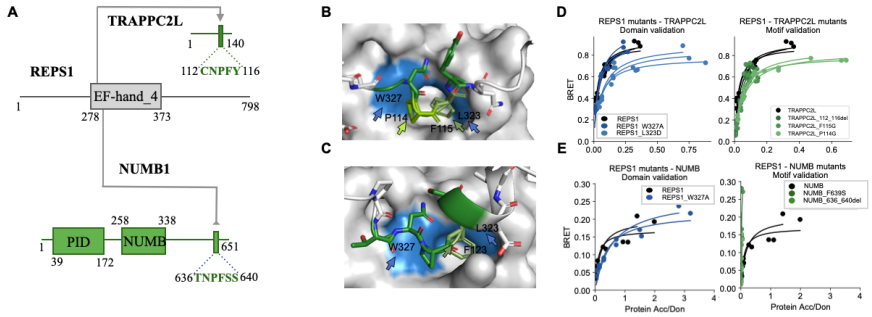


Figure 3.10: DMIs mediating REPS1 interactions (A) Schematic illustrating REPS1 and its partners and their interactions mediated by interfaces. The edge ending points towards the predicted motif, where the arrow implies predicted DMI, while the half-circle points to the known DMI and gray indicates interaction mediated by different interfaces, where the question means that this interface was predicted by AF-MM using fragmentation approach. (B)Predicted interface interaction structure of the predicted DMI, where the domain is in contact with the second motif in the REPS1-TRAPPC2L interaction. The structure highlights mutated residues on the domain (in blue) and on the motif (in green), with arrows pointing to these residues. (C) Predicted interface interaction structure of the known DMI, where the domain is in contact with the second motif in the REPS1-NUMB interaction. The structure highlights mutated residues on the domain (in blue) and on the motif (in green), with arrows pointing to these residues. (D) Experimental validation of putative DMI of REPS1-TRAPPC2L using BRET saturation assay. (E)Experimental validation of known DMI of REPS1-NUMB using BRET saturation assay.

REPS1- TRAPPC2L

It was predicted that EH domains of REPS1 bind to the LIG_EH_1 motif, 112-116 of TRAPPC2L. We had only one prediction and a high DMI match score of 0.883. AF-MM modeled NPF residues of the motif fitting in the deep pocket of the domain (**Figure 3.10 A**). although it is an interesting prediction and the motif is docked as seen in the known structure 2JXC (not shown), the confidence score was low 0.6 (**Figure 3.10 B**).

The L323D mutation on the domain of REPS1 deeper in the pocket slightly reduced the interaction, while W327A close to the contact with edge residues of the motif did not affect the interface. The mutants on the motif, P114G weakend the binding. However, F115G and deletion of the motif did not disrupt the binding (**Figure 3.10 D**). It would be interesting to employ the fragmentation approach and predict the novel interface.

To sum up, we could test 14 out of 20 selected DMIs across 12 PPIs. Among 14 tested DMIs, we confirmed both binding regions in 5 DMIs (CTBP1-DMRTB1, WWOX-CPSF6 (ii), WWOX-CPSF6 (iii), WWOX-DAZAP2, PPP3CA-FAM167A) and partially confirmed 1 DMI (SPOP-

RXRB) for its domain region only out of 7 predicted DMIs. Additionally, we re-confirmed 5 (CTBP1-TGIF1, CTBP1-IKZF1, WWOX-LITAF(i), WWOX-LITAF(ii), IQCB1-CALML3 (iii)) and the motif region for 1 DMI (IQCB1-CALM1 (iii)) out of 7 known DMIs. We also tested 7 negative controls mediated by different interfaces and showed that 5 PPIs (CTBP2-CTBP2, WWOX-HOXA1, WWOX-CSNK2B, WWOX-SNRPC, PPP3CA-PPP3R2) might bind through different interfaces, while other 2 PPIs (SPOP-MYD88, IQCB1-MNS1) showed are likely to be wrong and require further investigation to define the interface between these interactions.

3.4 The application of the strategy of the variant effect on PPIs

Interaction profile for variants falling in WWOX

As comparative interaction profiling is challenging due to the scarcity of pathogenic mutations on motifs and the difficulty in crystallizing disordered regions for functional studies, many reported mutations in databases like ClinVar are not functionally validated and rely on predictive tools like PolyPhen-2, which showed limitations in predicting variant effect (**Sahni et al. 2015**).

This makes the interpretation of PPI profiling uncertain. We propose a PPI-centric strategy that incorporates domain-motif interface (DMI) information that seems to be suitable to better prioritize and interpret variants, providing a clearer understanding of their impact on protein interactions and contribution to disease. To showcase the application of our strategy we characterized the variants selected for the study.

WWOX, a protein involved in neural development and cancer, was chosen to explore the impact of specific mutations on its interactions. Three mutants—two VUS and one pathogenic mutation—were successfully cloned and experimentally tested (**Figure 3.11 A**). The pathogenic mutation, E17K, was documented in ClinVar and found in patients with developmental and epileptic encephalopathy (DEE). However, there is no evidence revealing the pathogenicity. The prediction was done by PolyPhen-2. According to AF-MM predicted structures, this mutation on the interacting WW1 domain is not in contact with the motif, suggesting it should not disrupt interaction at this site (**Figure 3.11 B**). Indeed, experimental data confirmed that the BRET signal for the WWOX-LITAF interaction remained similar to the wild-type, indicating no significant impact on this PPI (**Figure 3.11 D (i)**).

E17K also did not affect the interactions between LITAF-CSNK2B, WWOX-HOXA1 and WWOX-SNRPC interactions (**Figure 3.11 D (iv-vi)**). The effect on interactions with DAZAP2 and CPSF6 could not be observed due to instability issues with the mutant during co-expression **see Appendix, Figure 5.5 B & C**, necessitating further investigation (**Figure 3.11 D (iii, V)**). Overall, this pathogenic mutation did not disrupt the interface, implying that its clinical impact may involve other processes or that the mutation is not pathogenic as stated in ClinVar.

The VUS variant E17D demonstrated a similar interaction profile to the pathogenic variant E17K, showing no significant impact on interactions with CPSF6 and DAZAP2, though it was not tested with SNRPC (**Figure 3.11 D (ii, iii, v)**).

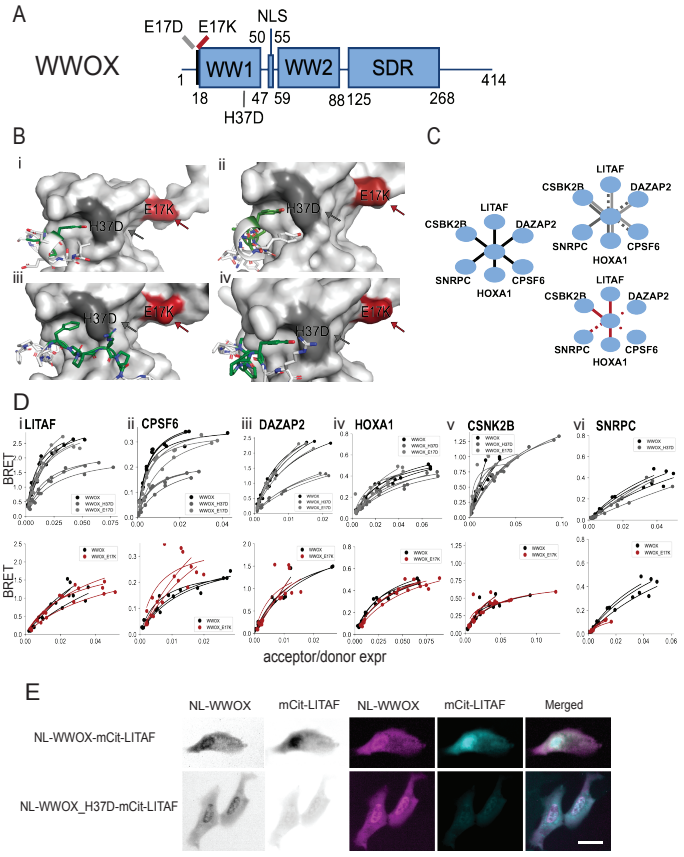


Figure 3.11: The effect of variants falling into the interface using interaction profiling. (A) Schematic illustration of the functional regions within WWOX with the location of variants, where the color indicates the pathogenic (red) and VUS (gray). (B) i Predicted interface interaction structure of the WW1 domain with the first PPSY motif in the WWOX-LITAF interaction. (B) iii Predicted structure illustrating the second motif on CPSF6 and tandem WW domains. (B) iii Predicted structure illustrating the third predicted motif on CPSF6 and tandem WW domains, where the motif is in contact with the second WW domain. (B) iv The putative model of the motif on DAZAP2 and tandem WW domains. (C) The schematic PPIs illustrate interaction profiles of wild-type and mutated interaction. (D) i Experimental assessment of the variants on known DMIs of WWOX-LITAF using BRET saturation assay. ii Experimental assessment of the variants on putative DMIs of WWOX-CPSF6 using BRET saturation assay. iii Experimental assessment of the variants on putative DMIs of WWOX-DAZAP2, iv on putative DMIs of WWOX-HOXA1, v on putative DMIs of WWOX-CSNK2B using BRET saturation assay. vi Experimental assessment of the variants on putative DMIs of WWOX-SNRPC using BRET saturation assay. (E) The microscopy experiment shows the localization of the H37D variant compared to the wild-type WWOX. The intensity of nanoLuc luciferase tagged LITAF (wild-type and mutant) was shown inverted and magenta and mCitrine tagged WWOX wild-type was shown inverted and cyan. scale = 10µm.

In contrast, the VUS variant H37D in WWOX found in patients with developmental and epileptic encephalopathy and autosomal reces-

sive spinocerebellar ataxia 12 is located within the DMI interface of DAZAP2, CPSF6, and LITAF. The substitution of histidine with a negatively charged aspartate could disrupt interactions by interfering with a tyrosine residue within the motifs (**Figure 3.11 B**). Experimental data confirmed this, as H37D notably disrupted interactions mediated by this interface (**Figure 3.11 D (i-iii)**), while interactions with proteins such as HOXA1, SNRPC, and CSNK2B remained unaffected (**Figure 3.11 D (iv-vi)**). These findings suggest that this variant H37D disrupts the interaction with partner LITAF, DAZAP2, and CPSF6 (**Figure 3.11 D (i-iii)**). WWOX binds to LITAF, a protein involved in mediating inflammatory responses and apoptosis. The ability of the WW1 domain to bind the motifs in these partners to regulate signaling processes. LITAF is critical for controlling inflammation and cell death. Such a disruption could hinder WWOX's regulatory role, leading to unchecked inflammatory responses or improper cell death signaling, potentially contributing to disease pathology.DAZAP2 is involved in RNA processing and signaling pathways that regulate cellular differentiation and proliferation. The interaction with WWOX may help modulate these pathways, ensuring an appropriate cellular response to stress and developmental cues. The disruption of this interaction might lead to altered RNA processing or dysregulation of signaling pathways, affecting cellular homeostasis and potentially contributing to developmental disorders. CPSF6 plays a crucial role in RNA cleavage and polyadenylation, processes essential for mRNA maturation. The binding of WWOX to CPSF6 could influence these processes by modulating RNA metabolism and gene expression regulation. The H37D variant could prevent proper binding of the WW1 domain to CPSF6, potentially affecting the function of the CPSF complex. This disruption might have widespread effects on gene expression, mRNA stability, and cellular response to DNA damage, which are critical in neurodevelopmental and neurodegenerative diseases.

This result suggests that profiling variants based on shared interaction disruption and DMI interface impact may be an informative approach to characterizing candidate disease-associated mutations. However, we cannot exclude the possibility that some expressed mutants might be partially misfolded or disrupt PPIs by altering protein compartmentalization. To test this, we used microscopy to verify whether mutant constructs alter localization compared to the wild-type protein (**Figure 3.11 E**). Our findings indicate that the H37D variant remains in the same cellular compartment as the wild-type WWOX, suggesting that the observed interaction disruptions are not likely due to mislocalization.

Interaction profiles of variants found in IQCB1

Mainly, mutations found in IQCB1 are associated with retinal disorders such as Senior-Loken syndrome 5 and Leber congenital amaurosis 10 (LCA 10). Many variants in IQCB1 are of uncertain significance, and the available evidence is currently insufficient to determine the definitive role of these variants in the disease.

For example, the R404G variant has been identified in patients with Nephronophthisis and other inborn genetic diseases and is classified as a Variant of Uncertain Significance (VUS). Algorithms developed to predict the effect of missense changes on protein structure and function, such as PolyPhen-2 ("Probably Damaging") do not consistently agree on the potential impact of this missense change. This variant has not been reported in the literature in individuals affected with IQCB1-related conditions and is not present in population databases (e.g., ExAC shows no frequency for this variant). The arginine residue (R404) is highly conserved and is predicted to be positioned deep within the domain of CALM1/CALML3 (**Figure 3.12 B (i)**). The change of the residue at this position might potentially affect the binding affinity and specificity, disrupting critical protein-protein interactions (PPIs). Another uncertain variant, N406Y (**Figure 3.12 A**), reported in ClinVar, also falls within the same motif and can disrupt PPIs similarly (**Figure 3.12 B (i)**). As expected, experimental data indicate a slight reduction in binding affinity for both interactions with CALM1 and CALML3 (**Figure 3.12 D (i)**).

Interestingly, the perturbing effect of these mutants was more pronounced when co-expressed with motif-binding CALML3 (**Figure 3.12 D (i, top right)**), suggesting a differential impact on binding efficiency between the two calmodulin-like proteins. This differential impact may reflect variations in the structural conformation or binding dynamics between CALM1 and CALML3, which could influence the pathophysiological consequences of these mutations.

IQCB1 is involved in several cellular processes, including cilia function and protein trafficking, which are crucial for maintaining photoreceptor cell integrity in the retina. The disruption of interactions with CALM1 and CALML3 due to mutations like R404G and N406Y could impair calmodulin-mediated signaling pathways, leading to defective cilia assembly or maintenance. This disruption might contribute to the pathology observed in retinal degenerative diseases such as Senior-Loken syndrome 5 and Leber congenital amaurosis 10. Moreover, altered calmodulin interactions could affect calcium homeostasis and cellular stress response, further exacerbating disease progression in affected in-

dividuals.

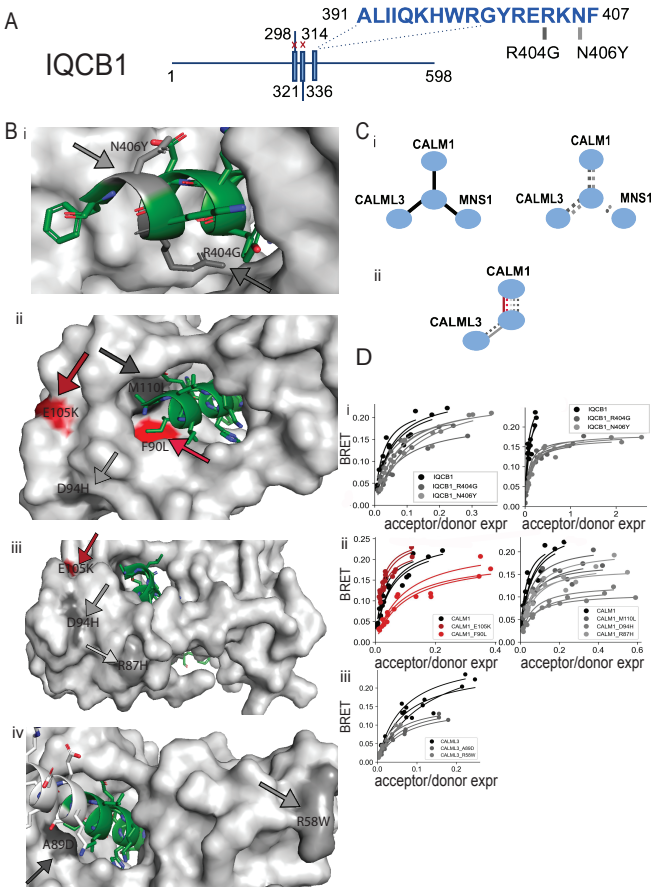


Figure 3.12: The effect of variants falling into the motif of IQCB1 using interaction profiling (A) Schematic illustration of the functional regions within IQCB1 with the location of variants, where the color indicates the VUS (gray). (B) i Predicted the interface interaction structure of the Eh domain of the domain in CALM1 in contact with the third motif. The structure shows the predicted interface and VUS variants (gray). ii Predicted structure illustrating the second motif on the Eh domain of the domain in CALM1 in contact with the third motif. The structure shows the predicted interface and pathogenic (red) with VUS variants (gray). iii The zoomed-out predicted structure is shown in ii. iv i Predicted the interface interaction structure of the Eh domain of the domain in CALML3 in contact with the third motif. The structure shows the predicted interface and VUS variants (gray). (C) i The schematic PPIs illustrate interaction profiles of wild-type and mutated IQCB1 interaction. ii The schematic PPIs illustrate interaction profiles of wild-type and mutated CALM1 and CALML3 interactions. (D) i Experimental assessment of the variants on known DMIs of IQCB1-CALM1 (left) and IQCB1-CALML3 (left) using BRET saturation assay. ii Experimental assessment of the CALM1 pathogenic (right) and VUS (left) variants on putative DMIs of IQCB1-CALM1 using BRET saturation assay. iii Experimental assessment of the CALML3 variants on putative DMIs of IQCB1-CALML3 using BRET saturation assay.

Calmodulin is an essential calcium-sensing, signal-transducing protein. Three calmodulin genes, CALM1, CALM2, and CALM3, have

unique nucleotide sequences but encode identical calmodulin proteins with 4 EF-hand calcium-binding domains. Calcium-induced activation of calmodulin regulates many calcium-dependent processes and modulates the function of cardiac ion channels. F90L is a pathogenic variant found in patients with LONG QT SYNDROME 14 and documented in Clinvar. The substitution occurs at a highly conserved residue between EF-hand domains II and III. The pathogenicity was not functionally studied. According to the position of the residue at the hydrophobic clutch of EF-hand domains (**Figure 3.12 B (ii)**), the experimental data showed that F90L disturbed the binding (**Figure 3.12 D (ii)**), while the other pathogenic variant E105K located outside of the interface (**Figure 3.12 C (ii)**) did not have any effect(**Figure 3.12 D (ii)**). This variant occurred de novo in a patient submitted for whole exome sequencing and it does not have functional evidence. Although the expression of this mutant is very high (**see Appendix, Figure 5.7 A**), it might partially destabilize the mutant causing the pathogenic effect. But it also might mean that the variant is not pathogenic, and the in-silico analysis reported in ClinVar is incorrect. Although E105 is outside the direct binding interface, it is part of the hydrophobic clutch that mediates the interaction between the EF-hand domains. A disruption here could impair the coordinated movement and proper orientation of these domains, reducing the ability of calmodulin to expose the necessary hydrophobic patches for binding target proteins effectively.

We also tested the three VUS variants, where M110L in CALM1 was predicted to be in the domain, and D94H and R87H were predicted to be outside of the interface (**Figure 3.12 D (iii, iv)**). Surprisingly, M110L caused a slight reduction, while the D94H variant was found in patients with Catecholaminergic polymorphic ventricular tachycardia 4 and Long QT syndrome 14 significantly affected the interaction with CALM1 (**Figure 3.12 D (ii)**). The VUS A89D in CALML3 also showed the effect on interaction with IQCB1 (**Figure 3.12 D (iii)**).

Interaction profile for variants falling in SPOP

We also tested pathogenic variants detected on the MATH domain of SPOP on the interaction with its partners RXRB and MYD88 (**Figure 3.13 B (i-ii)**). As expected the mutants perturbed the interaction with the predicted motif on RXRB (**Figure 3.13 D (i)**). Interestingly, the Y87C variant did not affect BRET with MYD88 (**Figure 3.13 D (ii)**). However, the predicted interface might not be correct, as it was previously shown in literature and in this study, it was expected that these mutants might not be on the correct interface with MYD88 and

have no effect on binding. In agreement with this assumption, the VUS variants also did not change the interaction with MYD88.

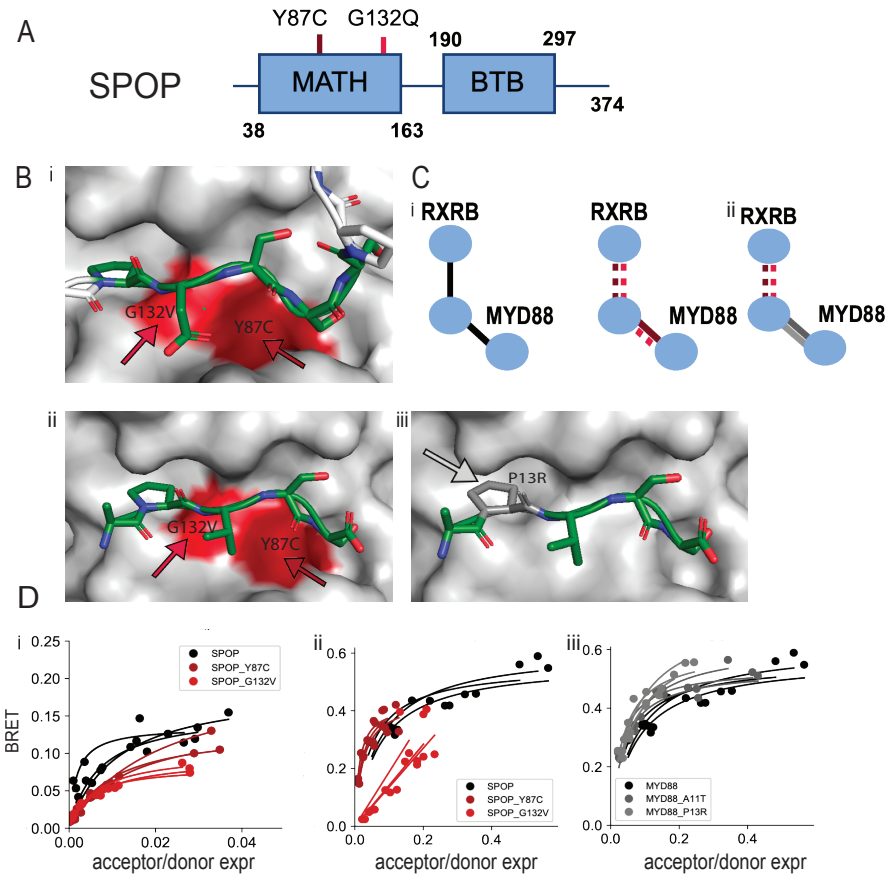


Figure 3.13: The effect of variants falling into the motif of SPOP using interaction profiling. (A) Schematic illustration of the functional regions within SPOP with the location of variants on MATH domain, where the color indicates the pathogenic (red) variants. (B) i Predicted the structure of the MATH domain of the domain in SPOP in contact with the motif of RXRB. ii Negative control interaction SPOP-MYD88 using the novel interface using fragmentation AF-MM approach. The predicted model shows the MATH domain predicted to bind to the N-terminal motif on MYD88 with the pathogenic variants on the MATH domain. iii The same structure with VUS variant on a predicted motif in MYD88. (D) i Experimental assessment of the variants on known DMIs of SPOP-RXRB using BRET saturation assay. ii Experimental assessment of the effect of pathogenic variants on SPOP on the interaction with MYD88. iii Experimental assessment of the effect of VUS variants on moti of MYD88 on the interaction with SPOP.

The effect of variants sitting on motif

In addition, we tested successfully cloned mutants on the motifs of partners of our candidate partners LITAF, IKZF1, DMRTB1, FAM167A, DAZAP2, CPSF6 and TRAPPC2L. VUS variants found close to the first PPSY and on the second PPSY of LITAF (**Figure 3.14 B (i-ii)**) do not disrupt the interactions with WWOX (**Figure 3.14 C (i-ii)**).The lack of effect observed with the mutants could be due to the nature of the substitution; it may not be significant enough to affect the binding affinity between the two proteins, thereby failing to cause a noticeable disruption in the interaction. Further experiments could help clarify the extent of this variant's impact on different protein partners.

Additionally, this interaction is maintained by two PPSY motifs and the WW-1 domain, which can compensate for the loss of a single contact point, masking the effect of certain variants. It would be interesting to test the perturbation effect of this VUS on interactions with other partners mediated by the same DMI but only one interface to determine if the variant disrupts those interactions or keeps them similarly intact. In addition, the BLI experiment showed that mutant Y61D was localized similarly to the wild-type.

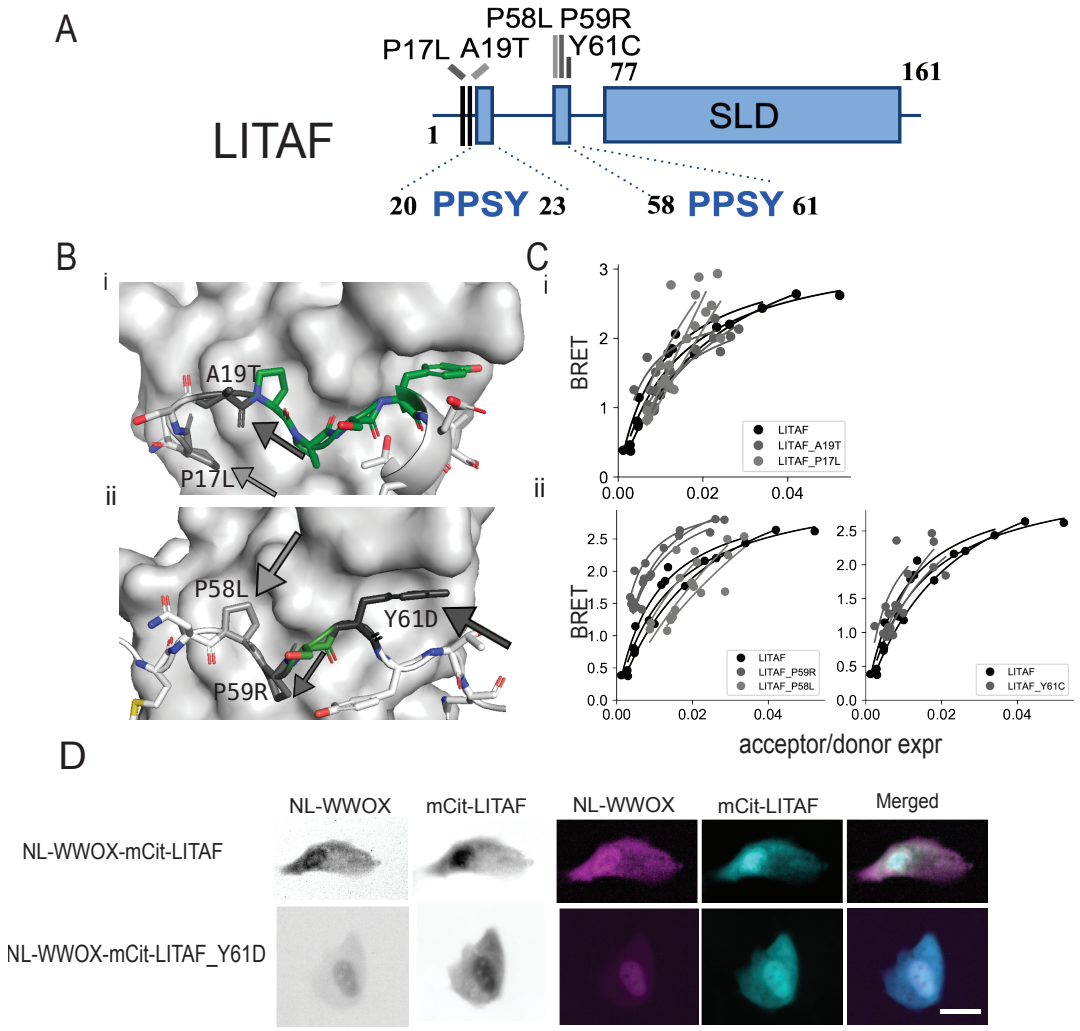


Figure 3.14: The effect of variants falling into the motif of LITAF using interaction profiling. (A) Schematic illustration of the functional regions within LITAF with the location of variants on motifs, where the color indicates the VUS variants. (B) i Predicted the structure of the WW-1 domain and recognized the first PPSY motif on LITAF. It also shows VUS variants situated close to the motif. ii Predicted the structure of the WW-1 domain and recognized the second PPSY motif on LITAF. It also shows VUS variants situated close to the motif. (C) i Experimental assessment of the VUS LITAF variants on known the first motif using BRET saturation assay. ii Experimental assessment of the VUS LITAF variants on known the second motif using BRET saturation assay. (D) The BLI experiment tested the localization of H37D variant compared to the wild-type WWOX. The intensity of nanoluc luciferase tagged LITAF (wild-type and mutant) was shown inverted and magenta and mCitrine tagged WWOX wild-type was shown inverted and cyan. The images of both interacting proteins are merged.

The variant located on the flanking regions close to the predicted motif of IKZF1 (M31V, S41L) showed a slight effect. VUS variants found on the motifs of FAM167A and DMRTB1, e.g. (R178H (**Figure 3.15 B (i-ii)**) and V8M (**Figure 3.15 C (ii-ii)**) showed lower BRET compared to wild-type (**Figure 3.15 B (iii) and C (iii)**)). On the other hand,

VUS located away from the motifs showed similar BRET results as wild-type interactions (**Figure 3.16**).

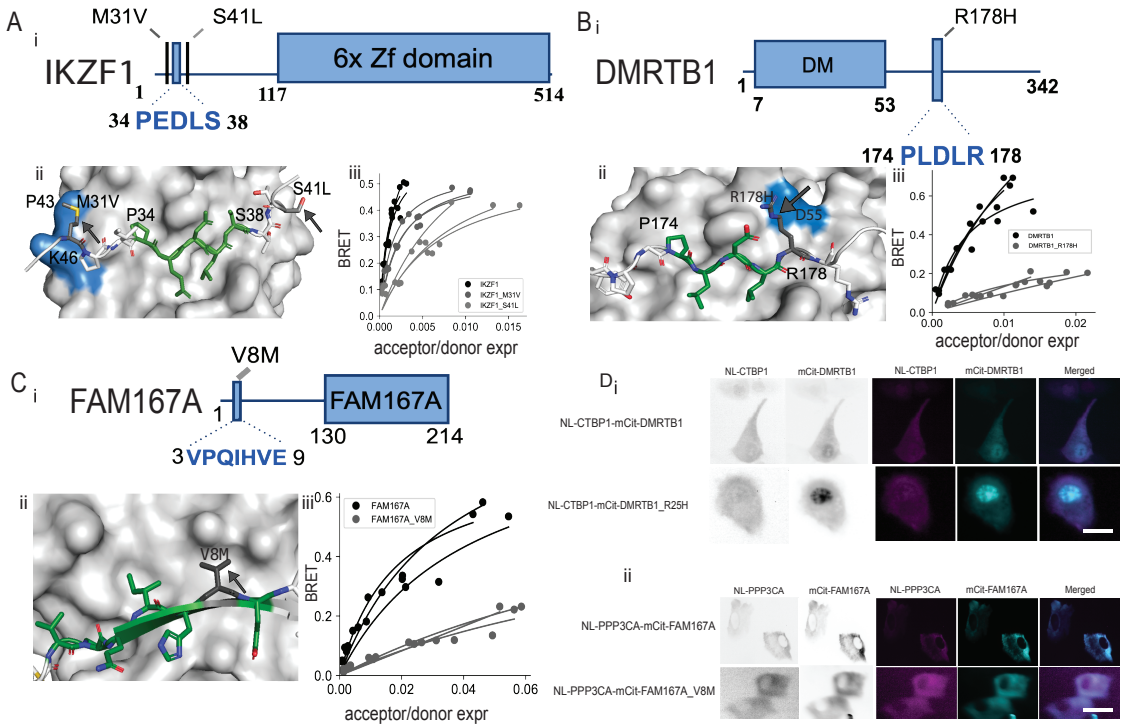


Figure 3.15: The effect of variants falling into the motif of IKZF1, DMRTB1 and FAM167A using interaction profiling. (A) i Schematic illustration of the functional regions within IKZF1 with the location of VUS (gray) variants on motifs. ii Predicted the structure of the CTBP1 domain and recognized the first PEDLS motif in IKZF1. It also shows VUS variants situated close to the motif. iii Experimental assessment of the VUS variants on a known motif in IKZF1 using BRET saturation assay. (B) i Schematic illustration of the functional regions within DMRTB1 with the location of VUS (gray) variant on the motif. ii Predicted the structure of the CTBP1 domain and recognized the first PLDLR motif on DMRTB1. It also shows VUS variant situated close to the motif. iii Experimental assessment of the VUS variant in putative motif in DMRTB1 using BRET saturation assay. (C) i Schematic illustration of the functional regions within FAM167A with the location of VUS (gray) variant on the motif. ii Predicted the structure of the CTBP1 domain and recognized the first PLDLR motif on DMRTB1. It also shows VUS variant situated close to the motif. iii Experimental assessment of the VUS variant on a putative motif in FAM167A using BRET saturation assay.

Here we evaluated the effect of variants on the PPIs mediated by domain-motif interfaces. We showed that variants located within the motif region can disrupt interactions, potentially altering the function of these interactions and contributing to disease development. Moreover, mutations near the motif region may also slightly affect the interface, potentially disrupting the biological processes mediated by these interactions.

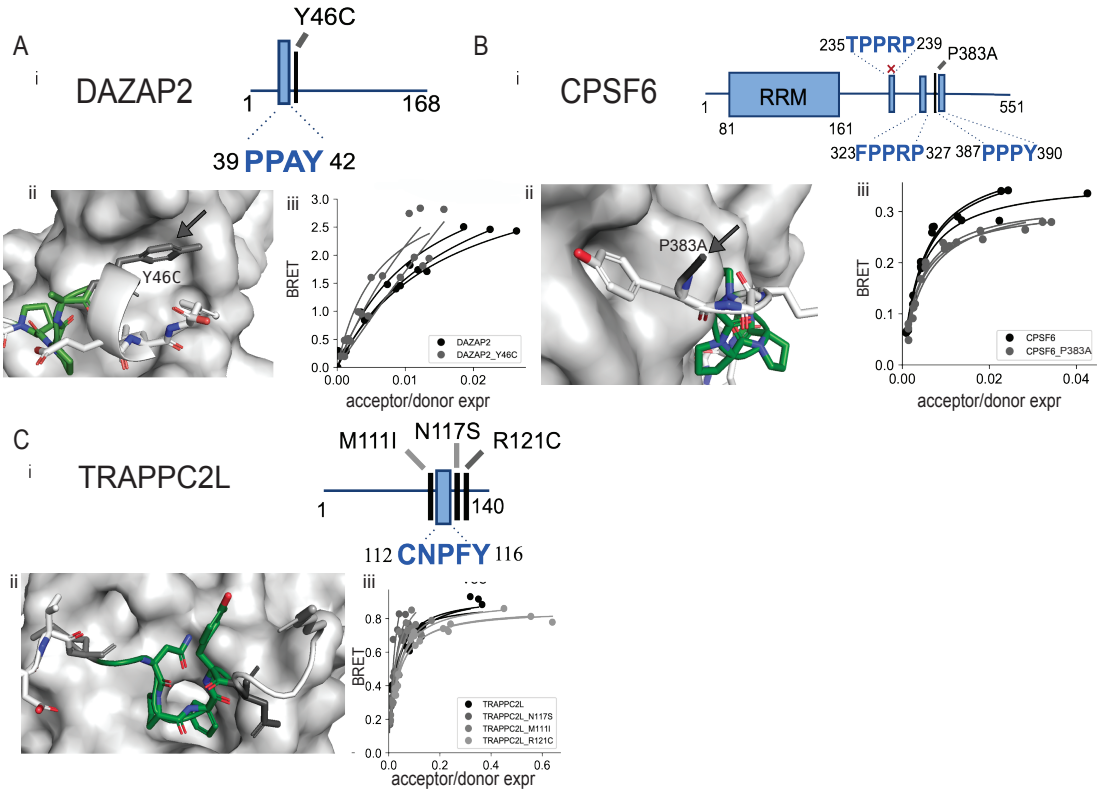


Figure 3.16: The effect of variants falling into the motif of DAZAP2, CPSF6 and TRAPPC2L using interaction profiling. (A) i Schematic illustration of the functional regions within DAZAP2 with the location of VUS (gray) variant close to the motif. ii Predicted the structure of the WW-1 domain and recognized the predicted motif in DAZAP2. It also shows VUS variant situated close to the motif. iii Experimental assessment of the VUS variant on a predicted motif in DAZAP2 using BRET saturation assay. (B) i Schematic illustration of the functional regions within CPSF6 with the location of VUS (gray) variant close to the third motif. ii Predicted structure of the WW-1 domain and recognized the third motif on CPSF6. It also shows VUS variant situated close to the motif. iii Experimental assessment of the VUS variant in putative motif in CPSF6 using BRET saturation assay. (C) i Schematic illustration of the functional regions within TRAPPC2L with the location of VUS (gray) variants close to the motif. ii Predicted the structure of the Eh domain in REPS1 and recognized putative motif in TRAPPC2L. It also shows VUS variant situated close to the motif. iii Experimental assessment of the VUS variants on a putative motif in TRAPPC2L using BRET saturation assay.

.However, not all mutations within the interface necessarily disrupt the interaction. Some residues, even when mutated, may not significantly alter the interface if the substitution does not substantially change the binding strength. Additionally, some interactions may be stabilized by multiple interfaces, which can compensate for the loss of a single contact point, masking the effect of certain variants (e.g. LITAF-WWOX).

On the other hand, the disruption of the interaction might be caused by partial folding or mislocalization of the mutant, rather than direct interference with the binding interface. Therefore, additional studies are

needed to confirm these possibilities and to determine whether observed disruptions are due to changes in protein structure or localization.

Overall, our findings indicate that integrating interaction disruption profiles with DMI interface information can enhance our understanding of variant effects in the context of PPI interactions.

This combined approach allows for a more nuanced characterization of variants, potentially leading to better identification of disease-associated mutations and providing deeper mechanistic insights into their role in disease pathology. However, considering the complexity and number of interfaces that mediate interactions, the diverse biological processes they influence, structural conformations and the specific properties of each amino acid at the contact sites, and the residue it is changed to this strategy can be further refined to achieve more accurate and controlled results.

Chapter 4

Conclusion and future perspectives

4.1 Deciphering protein interaction interfaces using DMI predictor tool

The development of the DMI tool and its application to HuRI annotate about 3200 protein-protein interactions (PPIs) with high-confidence putative DMI interfaces (**see Chapter 3 section 3.3.)**, providing valuable insights into the mechanistic functions of these interactions. This advancement has greatly enhanced our ability to understand how specific mutations might disrupt interactions, aiding in the characterization of variants found in patients. By analyzing how a variant perturbs PPIs, we can hypothesize its potential contribution to the development of disease symptoms or aetiology. Such hypotheses can then be tested through downstream experiments, which is crucial for the advancement of precision medicine.

Despite these advancements, there is still room for improvement in the performance of the DMI tool. One issue is its inability to distinguish between repetitive tandem domains, (e.g. RR1 and RR2), which often appear sequentially within proteins and may serve different functions in mediating interactions. Incorporating domain-specific annotations and functional classifications can help differentiate between tandem repeats by considering their unique roles and sequence patterns. Advanced pattern recognition methods and contextual analysis can refine sequence analysis.

Another limitation identified during manual analysis is that some predicted DMIs did not meet the cutoff due to low IUPred scores, despite the motifs being disordered. This issue is likely due to the window-based nature of IUPred, where regions adjacent to folded segments are often predicted as folded. Therefore, enhancing the window size or incorporating additional prediction tools could improve the tool's ability to detect

likely true motifs, for example, AF-MM.

Our findings also showed that the variants found on flanking regions of the motif can also slightly affect the interactions, suggesting the potential involvement of these regions in maintaining the interface of a PPI. This insight can be integrated into the refinement of the potentially functional regions and variant effect characterization, helping to refine the understanding of how flanking regions contribute to interface stability and potentially influencing the assessment of variant impacts on protein interactions (**Luck et al. 2012**).

Furthermore, with the recent update of the ELM database, which has enriched SLiM classes with new instances, re-running the DMI tool using this updated dataset could significantly enhance prediction accuracy and outcomes. In addition, our predicted and experimentally validated data can

4.2 The application of DDI predictor and AlphaFold to map the PPI data with interaction interfaces

There is an overwhelmingly large number of PPIs that are not mapped with any known interfaces pointing to the fact that many interface types remain still uncovered, especially those involving motifs (**Rolland et al. 2014**; **Tompa et al. 2014**).

To detect these interfaces, the AF-MM approach can be employed to identify novel interfaces, which can then be mapped onto PPIs and overlapped with mutation data, as demonstrated in Chapter 2, Article II. All in all, using AF-MM to discover novel interfaces holds great potential as it bypasses the need for a reference list of interface types for interface searching. However, scaling this approach for higher throughput will require further development.

Another type of interface that was not mapped in this study is a domain-domain interface (DDIs). Given the more stable nature of folded domains and the interactions they mediate, structural information on DDIs is more abundant compared to DMIs. For example, the 3did database extensively catalogs DDIs (**Geist et al. 2024**). Our lab has assessed the quality of DDIs in 3did providing us with important insights regarding features that can aid in scoring predicted DDIs for their abilities to mediate PPIs (**Geist et al. 2024**). Incorporating these insights can improve the mapping of the PPI dataset with DDIs that help to interpret the effect of variants on protein function.

4.3 Enhancing Predictive Accuracy of Variant Effects and Mutation Design through Positioning on Predicted AF-MM Interface Structures

The application of AF-MM to predict novel interfaces, for which the resolved structures are not available, significantly aided in understanding how well the putative motif fits into the binding pocket through visualization. This process allows for the detection of residues in close contact, the assessment of the structural location of mutated residues, and the design of mutants for experimental validation. While the structural information can give insight into the predicted interfaces and help in variant characterization, the manual inspection of predicted structures and the localization of variants is time-consuming. To address this limitation, my colleague is currently working on applying AF-MM to the entire set of DMIs with overlapping pathogenic variants of VUS mutations to analyze and implement the structural information.

4.4 Improvement of the BRET assay to validate the predicted interfaces

While the medium-throughput cloning pipeline and BRET assay developed in this study have been valuable for validating predicted interaction interfaces, several steps within the pipeline could be optimized to enhance both efficiency and accuracy. Currently, the manual picking of colonies is a bottleneck in the current plate-based medium-throughput pipeline, requiring substantial time and labor to select individual colonies for inoculation. Implementing automated colony pickers could address this issue by handling multiple colonies simultaneously with higher precision, thereby speeding up the workflow and reducing the risk of contamination or human error.

Using BRET assay we detected about 50% of protein-protein interactions from HuRI. Although this detection rate aligns with previous studies or was even higher, there is still room for improvement. Initially, we cloned fusion proteins exclusively at the N-terminal, based on the observations that the expression is better at the N-terminal. Trepe et al (2018). However, Trepte et al. have demonstrated that testing protein pairs in various configurations increases detection rates while maintaining low false detection rates (**Trepte et al. 2018**; **C. Trepte S. et al. 2021**). They also showed that tagging the proteins close to the interaction interface might improve PPI detection. While cloning tags

in different configurations or close to the interface could enhance the detection of PPIs, this approach might also increase the time required for cloning.

Additionally, choosing a more sensitive fusion tag can enhance the detection capability of the BRET assay regardless of the tag's position relative to the interaction interface. For example, using tags with higher quantum yields or those that offer better resonance energy transfer efficiencies can lead to stronger and more reliable BRET signals. For example, using mNeonGreen as an acceptor fluorophore in BRET assays significantly increased the dynamic range and sensitivity compared to traditional GFP derivatives (**Shaner et al. 2013**). These more sensitive tags could improve detection sensitivity even if the tag is not in the optimal position relative to the interaction interface.

The NL and mCit fusions used in the BRET assay allowed us to monitor the expression levels of wildtype and mutant constructs, which is important to rule out loss of binding because of a destabilization of the protein. However, we cannot exclude the possibility that some expressed mutants might still be partially unfolded or mislocalized and thus, some loss of binding detected in our study could be unspecific and not the result of a specific perturbation of the predicted interface (**Lacoste et al. 2023**).

Advanced imaging techniques could be scaled up and integrated into the workflow to assess whether mutant proteins are mislocalized. This approach would help determine if the observed binding loss is due to the mutant proteins being in the incorrect cellular compartment rather than a direct effect on the interaction interface. While I attempted to implement BRET-based bioluminescence imaging (BLI) to test whether the localization of a mutant is not changed compared to the wild-type protein, we faced challenges in the setup of experimental steps that needed to be optimized for robust quantitative analysis. This optimization involves the finding optimal amount of cells for seeding, the DNA ratio for more efficient transfection, the concentration of the transfection agent as well as downstream analysis involving defining regions of interest (ROIs) for specific cellular compartments, using either manual methods or automated segmentation algorithms. This approach allows for assessing whether the mutant proteins overlap with these markers and determining any shifts in localization. Further statistical analysis would ensure that any observed differences are significant and not due to random variation. Implementing these strategies will help confirm if localization changes contribute to the observed effects, thereby providing a more accurate interpretation of the impact of mutations on protein function.

Along with mislocalization studies, BRET-based imaging can be used for the detection of a single BRET within the cell (**Dragulescu-Andrasi et al. 2011**; **Kobayashi et al. 2019**). The determination of BRET per cell might enhance the precision of interaction studies by providing detailed insights into how individual cells contribute to the overall interaction dynamics. Quantifying BRET signals per cell allows for a more granular analysis of the interactions, potentially revealing variations in PPI strength and localization that may be masked in bulk measurements. Moreover, BRET-based microscopy can be applied not only in mammalian cells but also in tissues and in vivo animal models (**Dragulescu-Andrasi et al. 2011**). Kobayashi et al. (2019) demonstrated the use of BRET-based imaging to monitor protein interactions and subcellular localization in live animal tissues. Their study emphasized that BRET, with its enhanced dynamic window due to reduced background signals, is particularly effective for detecting subtle changes in protein interactions. They also illustrated the quantification of BRET signals, including the dissociation of protein complexes and redistribution within cellular compartments. For instance, they used manual segmentation and pixel-by-pixel analysis to quantify BRET signals from specific subcellular regions, revealing significant changes in protein interactions upon receptor activation. This quantitative approach enabled precise measurements of BRET signal changes, facilitating detailed insights into dynamic biological processes such as receptor endocytosis and protein localization in vivo. However, it was done on a small scale. The development of a plate-format scalable BRET-based BLI pipeline has to be addressed.

4.5 General outlook

This thesis has proposed a strategy driven by prediction and experimental validation of domain-motif interfaces and integrating this information to interpret the effect of uncharacterized variants on protein function. In doing so, we have gained profound insights into the intricate interplay between different functional modules, such as domains and motifs in proteins that facilitate their interactions. Moreover, using this strategy we provided experimental evidence and structural information on the effect of variants falling into DMIs mediating protein-protein interactions. This information can be explored in future studies aimed at delineating potential molecular mechanisms causing disease.

Given the useful mechanistic insights that prediction tools like the DMI predictor tool can provide, I expect the optimization and applica-

tion of these tools (DDI predictor and AF-MM) in mapping PPI with interfaces to bring us closer to a fully structurally annotated human protein interactome mapped with interfaces. Moreover, I anticipate greater inclusion of interface information in experimental workflows, where this will help generate hypotheses to guide experiments and aid in variant characterization.

Binary interaction assays like BRET have proven to be suitable tools for validating PPI interfaces, but there are still several ways to further enhance their capabilities in characterizing variant effects on PPIs(**Dragulescu-Andrasi et al. 2011**; **Kobayashi et al. 2019**). In addition to expanding the power to systematically assess the effects of variants on protein-protein interactions (PPIs), it is crucial to implement systematic downstream steps (e.g., reporter assays, cell proliferation, apoptosis assays) to gain deeper insights into how these variants impact biological processes. By integrating these additional steps, researchers can move beyond just identifying whether a variant disrupts a specific interaction and start understanding the functional consequences of these disruptions within the context of cellular pathways and networks. I anticipate seeing the advancements of this assay in this direction.

Chapter 5

Appendix

5.1 Protocols

5.1.1 The medium-throughput cloning protocol

Medium-throughput GATEWAY cloning protocol

Data organization in MySQL DB cloning_data
Every HTP cloning project should have a bioinformatician assigned to it who helps with putting the data in the tables. Everything that the experimentalist can do on his/her own should not be done by the bioinformatician.

table project_descr:

column_name	content
project_id	e.g. CL01
experimenter	e.g. Christian
bioinformatician	e.g. E[illegible]
descr	e.g. cloning project for XL-MS project, more PRS pairs, every ORF cloned in N-ter NL and N-ter mCit vector
date_started	e.g. 2022-10-11

table orf_pairs:

column_name	content
project_id	e.g. CL01
orf_a	e.g. 49583 → if the ORFs of a pair are in no particular order -> put smaller orf_id as orf_a, otherwise describe order of ORFs in descr column in project_descr table
orf_b	e.g. 98584

table entry_clone_info:

column_name	content
orf_id	e.g. 49583
orf_len_nt	e.g. 1980
entry_plate_id	e.g. GDEh81001 → plate name from ORFeome
entry_well_id	e.g. A01 → well ID from plate from ORFeome
entry_inoc_plate_id	e.g. CL01GEh_01
entry_inoc_well_id	e.g. C10
pcr_amplicon	0 or 1 → 1 if PCR product looks good on gel, 0 otherwise
comments	space to leave additional comments for an ORF if needed

table expr_clone_info:

column_name	content
orf_id	e.g. 49583
expr_plasmid_id	e.g. KL_11
expr_plasmid_name	e.g. pcDNA3.1 cmyc-NL-GW
LR_plate_id	e.g. CL01LR_01.1
LR_plate_well	e.g. C10
colonies	0 or 1 → 1 if there were colonies selected for picking, 0 otherwise
expr_plate_id	e.g. CL01GExh_01.1
expr_plate_well	e.g. B05
MP_elu_plate_id	this and the next 3 columns only need to be filled if rearray occurred
MP_elu_plate_well	
expr_dil_plate_id	
expr_dil_plate_well	
DNA_conc_ng_ul	e.g. 235
theor_DNA_conc_dil	e.g. 100
seq_confirmed_bb_fw	0 or 1
seq_confirmed_bb_rv	0 or 1
seq_confirmed_full_length	0 or 1
comments	space to leave additional comments for an ORF if needed

Material List:

In separate excel sheet with calculator for amounts

- you can find this checklist on the:
  group drive→ HTP_cloning_data→ templates→ CL00_checklist_consumables

Prior to start of cloning

Computational part

- 2 weeks in advance: get cloning project ID by checking on group drive in HTP_cloning_data folder what the last cloning ID was, increment by 1 count, i.e. if last cloning ID was Cl01 → make Cl02
- 2 weeks in advance: design and discuss with K[illegible]d bioinformatician for cloning project plate layout for ORF inoculation plates for Day 1 and plate layout for inoculation plates of picked LR transformants for Day 4 (the way the plates should be organized, code can be written but the rearray is only at day 4 possible)
  - consider to leave a well free for the water control for the PCR and if and which controls you would like to have for LR
- Design labels for your plates → to see an example for how plate labels should be designed and for which plates labels are needed → take a look at the file HTP_cloning_plate_labels_schematic.pdf → then start from the plate labels from a previous cloning project by making a copy of the plate_labels.txt file from a previous cloning project into the folder of your new cloning project on the group drive and modify accordingly → if you are not sure, PLEASE, ask → if plate labels are wrong, computational as well as experimental steps of your cloning project can go wrong
- Print the plate labels with the help of M[illegible]en (a template for the labels can be found on group drive, HTP_cloning_data, templates, HTP_plate_labels; please also read the explanation how to print these labels)
- Get plate layouts with the help of a bioinformatician → the scripts to design the plate layouts ideally need information about the plate labels

Experimental part:

- About 4 weeks in advance prepare competent DH5alpha E Coli cells → talk to M[illegible]en, the prepar[illegible] itself takes 1 week
- 2-4 weeks in advance do maxi, midi or miniprep of empty expression vectors and plasmids for LuTHy assay (KL_01, KL_02, KL_03, KL_06, KL_07, KL_11, KL_247)
- At least 2 weeks in advance, make a copy of the excel sheet with the list of reagents and save it in your cloning folder on the group drive, calculate your amounts and check that everything is available
- 2 weeks in advance check the amount of
  - PCR plates (order no.: 781352 from Brand)
  - PCR foil
  - Reservoir for LB medium (order no.: HT69.1 from Carl Roth)
  - Costar plates (order no.: 3799 from Corning)
  - Microplate aluminum sealing tape (order no.: 6570 from Corning)

        - Adhesive gas permeable seals (order no.: AB-0718 from Thermo Scientific)
        - Combitip advanced 1ml (order no.: 0030089.430 from Eppendorf)
        - Qtray with lid and divider (square plates for Agar; order no.: MLDVX6029 from VWR international GmbH)
        - E-Gel 96 1%Agarose (GP) (check the expiring date; order no.: G700801 from Invitrogen)
        - E-Gel 96 High range DNA marker (order no.: 12352019 from Invitrogen)
        - Steril/autoclaved 2ml Deepwell plates, 96 round wells (order no.: E2896-2110 from Starlab)
        - Qiaprep 96 Plus MiniPrep Kit (order no.: 27291 from Qiagen)
        - QIAvac 96 (vacuum system needed for the MiniPrep; order no.: 19504)
        - 1250 µL (blue) integra grip tips for a digital multichannel pipette
        - 125 µL (yellow) integra grip tips for a digital multichannel pipette
        - 12,5 µL (pink) integra grip tips for a digital multichannel pipette
- 1 week in advance - get familiar with
    - The digital multichannel pipette
    - All other equipment you will need
    - The excel sheet to calculate the amounts
    - SQL database
    - LuTHy assay transfection template
    - The scripts for the different steps
- 1 week in advance take all needed consumables on your bench or -20°C
- 1 week in advance - check the amount of:
    - 40% glycerol (sterile)
    - Proteinase K (2µg/µl)
    - HF PCR polymerase and buffer (from Protein Production CF)
    - dNTPs (NEB freezer at IMB)
    - 96 gel loading buffer (Homemade, recipe:10mM Tris-HCl, 1mM EDTA, 0,005% bromophenol blue)
    - LR clonase (from Protein Production CF, should be stored at -80°C or better -150°C)
    - SOC medium (~8ml/plate)
    - Needed antibiotic
        - Ampicillin (100mg/ml)
        - Kanamycin (30mg/ml)
        - Spectinomycin (50mg/ml)
    - LB medium
    - LB-Agar (250ml/square plate)
    - Sterile glass plating beads
    - Sterile toothpicks
    - SOC medium
- At least 1 week in advance, order sequencing barcodes for the plates (Starseq)
- Between 1 and max up to 5 days in advance, prepare square plates with agar
- At least 1 day in advance, sort ORFeome plates in new rack
    - Do this step with one additional person as helper
    - work with blocks of 7 plates, because they fit as one block in the rack
    - Presort your ORF plates into a new rack; you will need ~1h per 10 plates (including time to let the -80 come back to temperature): Take out rack from -80 freezer, close freezer, sort out the plates needed in a box with dry ice. Put the rack back in the freezer and sort the plates into a new rack according to the order you will pick from them. Let the freezer get back to -80˚C before you go for the next batch of plates.
- fill PCR protocol for X-reactions with calculations
- Stock SOC medium
- Cut small pieces of Alu foil for resealing of plates
- Aliquot expression vectors in PCR stripes
- Dilute primer for PCR in 1,5ml eppi
- Dilute and aliquot primer for sequencing in PCR stripes
    - after LR we are sequencing with forward and reverse primer at the same time
    - after running the sequencing pipeline you will see for which ORF you need to design primers for full length sequencing
    - Forward primer:

    - For N-terminal NL-fusion: primer #44 NanoLuc-398fwd (GAACGGCAACAAAATTATCGAC)
    - For N-terminal mCit-fusion: primer #47 mCitrine-547fwd (AGCAGAATACGCCCATCG)
- Reverse primer:
    - If there is no C-terminal fusion: primer #51 pEXP_rev (GGCAACTAGAAGGCACAGTC)

Overview of the plates

Step	Plate label (example)	Type of plate
Inoculation plate	CL01GDEh_01	Costar plate (#3799)
PCR plate	CL01PCR_01	PCR plate (#781352)
Gel plate	CL01Gel_01	PCR plate (#781352)
LR plate	CL01LR_01.1*, CL01LR_1.2*	PCR plate (#781352)
Transformation plate	CL01TR_01.1*, CL01TR_1.2*	PCR plate (#781352)
Agar plate	CL01TR_01.1a / CL01TR_01.1b, CL01TR_01.2a / CL01TR_01.2b	Qtray (#MLDVX6029)
Deepwell inoc plate	CL01GExDW_01.1a / CL01GExDW_01.1b CL01GExDW_01.2a / CL01GExDW_01.2b	Deepwell plate (#E2896-2110)
Glycerolstock plate	CL01GEx_01.1, CL01GEx_01.2	Costar plate (#3799)
MiniPrep elution	CL01GExMP_01.1, CL01GExMP_01.2	Costar plate (#3799)
DNA dilution plate	CL01GExDil_01.1, CL01GExDil_01.2	PCR plate (#781352)
DNA Database plate	CL01GExSt_01.1, CL01GExSt_01.2	PCR plate (#781352)
DNA sequencing plate (forward and reverse)	CL01GExSF_01.1, CL01GExSF_1.2 CL01GExSR_01.1, CL01GExSR_1.2	PCR plate (#781352)

All plates should be labeled at the left side (having A1 top left corner)
*where applicable plates labelled x.1 contain NL fusion constructs, plates labelled x.2 contain mCit fusion constructs

Day 1 Picking and inoculation of ORFs (~1.5h just picking)

Checklist:

- 70% EtOH and tissues - to sterilize the plates from the Orfeome collection
- 50 mL falcon tube - to prepare mix of LB medium with corresponding antibiotic
- Tips - for picking up the ORF from the collection plate
- Alu foil cut in small pieces (size of one well) to close the opened wells with ORF
- 50 mL serological pipette
- Pipette boy
- 100 µL pipette
- Pipette tips
- 96-well, costar plate (Corning,#3799) - for the inoculation of the ORFs
- Adhesive gas permeable seals (order no.: AB-0718 from Thermo Scientific)
- Multichannel pipette with tips
- LB medium (200 µL per well)
- Antibiotic (Kanamycin or Streptomycin - 0,2µl/well)
- Dry ice box - for keeping the plates from the ORFeome collection while picking the ORFs

Do the following steps with one, better two additional people as helpers!

steps:

1. Use aseptic bench working technique
2. Label the inoculation plate (CL01GDEh_01)
3. Prepare a master mix of LB medium and antibiotic in a 50ml Falcon and vortex
    a. 200µl LB medium/well
    b. Pay attention to which ORF needs which antibiotics
    c. 1:1000 mixing of antibiotic to LB medium (i.e. 1µL into 1000µL of LB medium)
4. Prepare the reservoir for the LB medium
5. Pour the antibiotic LB mix in the reservoir
6. Use the 300µl multichannel pipette to distribute 200µL of antibiotic LB mix to each well in the 96-well plate
7. Take out the box with dry ice and put the first 7 plates with the needed ORFs on it
    a. make small stacks to keep cold
8. **Work as fast as possible on dry ice here**
   (working with 3 people simplifies the process, 1. person is taking the plates out, 2. person is picking the ORFs, 3. person is controlling)
    a. Disinfect the alu foil of the orfeome plate
    b. With the tip/toothpick make hole in the selected well
    c. Take another tip and scratch the ORF
    d. Then put it into the well of the inoculation plate with LB/Antibiotic medium, stir for a few seconds and discard the tip
    e. Immediately close the hole with the pre-cut alu foil pieces
9. Repeat steps 8b - 8e for each well to pick from a plate
10. Move to the next plate until done with the first batch then go back to step 7
11. Seal the inoculation plate with the air permeable adhesive seal
12. Incubate the plate overnight at 37 ˚C, 190rpm,
    a. cover with a paper box (to reduce evaporation)
    b. Incubator in the Niehrs lab

Day 2 PCR, Glycerolstock, E-Gel and LR reaction

Checklist PCR:

- 96-well skirted PCR plate - for PCR reactions
- 50 mL falcon tube - to prepare PCR master mix, 50ml because of multipette
- Alu foil - to cover the glycerolstock plate
- PCR foil to cover the PCR plate
- Multichannel pipette
- Multipette and combitips 1ml
- 100 µL (yellow) integra pipette tips special for a digital multichannel pipette
- 10 µL (pink) integra pipette tips special for a digital multichannel pipette
- 10ml reservoir - to pipette the reaction and transfer to the PCR plate
- Ice block - to keep PCR components in cold
- 40% glycerol stock (= 10 mL)
- PCR components
- PCR plate containing inoculated ORFs - for PCR
- E-Gel

Checklist for E-Gel:

- PCR plate
- Compitip advanced 2,5ml
- 96-well E-Gel
- 96 gel loading buffer (Homemade, recipe:10mM Tris-HCl, 1mM EDTA, 0,005% bromophenol blue)
- DNA marker E-Gel
- 50 µL manual multichannel pipette for loading the gel
- 200 µL tips
- BioRad Detection Machine

Checklist LR reaction:

- Cold PCR block: thermomix block keep it in the cold
- 2 PCR plates for mCit and NL fusions
- PCR plate (**CL01PCR_01**) with PCR products
- DNA for expression vectors (KL_11 & KL_247), diluted to 200 ng/µl, aliquoted to PCR stripes with 20µl each
- Ice box
- Autoclaved water
- 12.5 µL multichannel pipette (Tick)
- 125 µL multichannel pipette (Trick)
- 100 µL (yellow) integra pipette tips special for a digital multichannel pipette
- 10 µL (pink) integra pipette tips special for a digital multichannel pipette
- Rack for eppi tubes for LR clonase (the distance of the small racks work with the digital multichannel pipette)

PCR
PCR program:

Temperature	Time	Repeat	Step
98˚C	30s	1x	Initial denaturation
98˚C	10s	30x	Denaturation
55˚C	30s	30x	Primer annealing
72˚C	3min	30x	Extension
72˚C	5min	1x	Final extension
16°C	∞	1x	Hold

Master Mix

PCR components	Per 1 reaction (=1 well)	Per 100 reactions
Primer #48 pENT-F 10µM	2.5 µL	250 µL
Primer #49 pENT-R 10µM	2.5 µL	250 µL
dNTPs (10mM each dNTP)	1 µL	100 µL
10x High fidelity polymerase buffer	5 µL	500 µL
High fidelity DNA polymerase	0.55 µL	55 µL
H_2O	34.45 µL	3445 µL (= 5x 689 µL)

Steps PCR:

1. Label the PCR plate (CL01PCR_01)
2. Once the PCR components start to thaw. Vortex each PCR reagent
3. Prepare a master mix of all PCR components (see table)
    a. In 50mL falcon tube
4. Pipette 46 µl of the master mix in each well of the PCR plate
    a. Using the multipette and combitip 1ml (on ice/cold block)
5. One well should be used as control (master mix without ORF)
6. Remove the airpore seal of the inoculation plate
7. Close the inoculation plate with aluminum foil
8. Vortex the inoculation plate
9. Carefully remove aluminum foil
10. Transfer 4µL of the inoculated ORF culture to the PCR plate
    a. With the manual 10µL multichannel pipette
    b. The ORF layout is the same
    c. Always use new tips
11. Close the PCR plate with PCR foil
12. Vortex the plate briefly
13. Centrifuge briefly

14. Run the PCR (~3 hours)

Steps glycerolstock:

1. Check two wells how much bacteria culture is left
2. Then remove a “certain” amount to have 100µl of bacteria culture left in the inoculation plate
3. Add 100 µL of sterile 40% glycerol to each well of the inoculation plate (1:1 ratio)
4. Close the plate with alu foil
5. shake 45 sec at 800 rpm on the thermomixer
6. Store at -80°C (rack 8)

Validation of the PCR product with E-gel

- Info:
    - PCR products can be stored at 4°C for 48h, for longer time freeze PCR products
    - Document all wells that do not look ok on gel -> this info needs to go into MySQL table, send info to bioinformatician

Steps:

1. Label the E-Gel plate (i.e. CL01Gel_01)
2. Pipette 25 µl of blue 96 gel loading buffer in the E-Gel plate
    a. Using the multipette and 2,5ml Combitip
    b. Can be done while PCR is running
3. Add 6 µl of PCR product to each well
    a. Using the 10µl multichannel pipette
4. Install 96 well E-gel to the motherbase
5. Load 20µl PCR/buffer mix to each well
    a. Using the 50µl multichannel pipette
6. Load 20µl of E-Gel 96 High range DNA marker
7. All empty wells must also be filled with 20µl
    a. With buffer or loading dye
8. Insert the plug into the socket
9. Run gel for 12 min
    a. Program EG
10. Take picture with GelDoc Station
11. Analyze gel picture with the E-Editor 2.0 software
    a. On the desktop PC in the technical room
    b. Realign the bands and save it in your cloning project folder
    c. The software is pretty self-explanatory and has a manual available under the help button. Ask [illegible] elp.
12. Decide if PCR was successful and whether it is worth proceeding
13. Document all wells that did not look ok
    a. Add this information to the MySQL table entry_clone_info

LR reaction

Components	Per 1 reaction (1 well)
H2O	5,5 µL
Destination vector (200ng/µl)	1 µL
PCR product	1 µL
4x LR clonase	2,5 µL

1. Label the LR plates (CL01LR_01.1, CL01LR_1.2)
2. Decide if you want to include controls for the LR reaction
    a. I.e. no clone (only water), no LR clonase
3. Take out the destination vectors and put it on the bench to thaw
    a. Prepare a PCR stripe with 8x 20µl of KL_11 (NL-GW)
    b. Prepare a PCR stripe with 8x 20 µl KL_247 (mCit-His3C-GW)
4. In a clean reservoir pour ~ 2 mL of autoclaved water
5. Add 5,5µl water into each well
    a. Use 125 µL multichannel pipette (Trick)
    b. Aspirate 66 µl water and distribute 12x 5.5µl
    c. Repeat for the second plate.
6. Add 1µl of PCR product
    a. Use 12.5 µL multichannel pipette (Tick)
    b. Aspirate 2 µL of PCR products
    c. load 1 µl to each PCR plate for LR reaction.
7. Add 1µl of the NL expression vector KL_11
    a. Into the plates, which will contain NL fusions in the end (i.e. CL01LR_01.1)
    b. Use multichannel pipette
8. Add 1 µL of the mCit expression vector KL_247
    a. Into the plates, which will contain mCit fusions in the end (i.e. CL01LR_01.2)
    b. Use multichannel pipette
9. Take 8 tubes 4x LR clonase out and put on a rack
    a. **Info:** if the LR clonase is still very cold - it is difficult to pipette, LR clonase will be outside of the tip and the resuspension step gets difficult
    b. better: for 1 plate you will need 8 tubes of LR clonase (each tube contains 40µl) - vortex the tubes, centrifuge LR clonase, wait until LR clonase is easy to pipette (~2min) and then start, the leftover of the LR clonase should be discarded
10. Vortex each LR clonase twice for 2 seconds and put back to the rack
11. Add 2.5 µL LR clonase
    a. Use 12.5 µL multichannel pipette (Tick)
    b. Program HTP_LR
    c. Resuspend and discard the tips
12. Repeat until all wells of both plates has received LR clonase
13. Cover LR-plates with alu foil
14. Incubate overnight at 25°C (in PCR machine)
15. Close the plate with the PCR reaction with alu foil and store the PCR products at -20°C

alternative

1. Prepare all needed components for LR
2. Prepare a master mix of your expression vector and water (can be done several days before)
3. Aliquot 6,5µl of water/expression vector mix into both LR plates
    a. with the multipette
4. Add 1µl of PCR product
    a. with 10µl multichannel pipette
5. Take 8 tubes of 4xLR clonase
6. Vortex each LR clonase twice for 2 seconds and put back to a rack
7. Add 2.5 µL LR clonase
    a. Use 12.5 µL multichannel pipette (Tick)
    b. Program HTP_LR
    c. Resuspend and discard the tips
8. Repeat until each well of both plates has received LR clonase.
9. Cover LR plates with alu foil
10. Incubate LR plates overnight at 25°C (PCR machine, Thermoblock)
11. Close the plate with the PCR reaction with alu foil and store the PCR at -20°C

Stop point: LR plates could be stored at -20°C until processing with transformation

Preparing square agar plate (should be done at least the day before needed)

Check list

- LB Agar (250ml/square plate)
- Square plates and divider
- Microwave

1. Take LB-Agar (250ml) from IMB media lab
2. Use aseptic bench working technique
3. Heat Agar in the microwave (program: soften/melt, 2= melt dark chocolate, 100 = 5,5 min; after 3x the agar is liquid)
4. Let it cool down (i.e. add a clean stirrer to the agar and place the bottle on the magnetic stirrer, adjust the temperature to 50°C and 250rpm)
5. Add antibiotic (250µl) when the agar is cooled down sufficiently and you are ready to pour the plates
6. Take out the plate from the plastic protection
7. Add agar to the plate (pop bubbles with a pipette tip or move them to the side)
8. Take out the grid from the plastic protection
9. Add the grid in the square plate with agar
   --> the grid does not stay down - weigh down the grid with something (i.e. a 250ml bottle)
10. Let the agar solidify
11. Store at 4°C (upside down)

Day 3 Proteinase K digestion, Transformation and plating

Check list

- Proteinase K (2µg/µl)
- Ice box for 2 PCR plates with LR reaction
- Ice box for 2 PCR plates with competent DH5α cells
- Thermoblock/PCR machine for heat shock
- Thermoblock/PCR machine for 2 plates for recovery step
- 2 racks for 2 PCR plates
- 10ml reservoir to pour SOC medium
- SOC medium (8ml/plate)
- DH5α (2 PCR plates with aliquots of 30 µL)
- Square plates with agar (48 wells, 4 plates needed for 1 inoculation plate)

Proteinase K digestion and Transformation: (~3h)

1. Take out SOC medium (for one well = 80 µL, for 1 plate = 8 mL) and let it thaw at room temperature
   a. 50ml takes long time to thaw, could be placed at 4°C the afternoon before
2. Use aseptic bench working technique
3. Take out DH5α from -80°C
   a. Put them immediately on the ice
   b. Let them thaw
   c. Label the plate (i.e. CL01TR_01.1, CL01TR_1.2)
4. Take out the LR plates from the incubation
5. Centrifuge briefly the LR plates
6. Add 1µl of Proteinase K into all wells
   a. Take out 8 tubes

    b. Use multichannel pipette
7. Vortex briefly
8. centrifuge briefly
9. Incubate at 37°C for 10min
10. Transfer the plates on ice
11. Transfer 10µL of each LR reaction into the DH5a plate
    a. Use a multichannel pipette
    b. Difficult to get the whole 10µl out (~7µl)
    c. No resuspension, no vortex when adding the LR reaction into the DH5a
    d. Close the plate with alu foil
12. Incubate for 30 minutes on ice (bacteria with LR product)
13. **Meanwhile**: set the thermoblock to 42°C for the heat shock and set thermoblock for 2 plates to 37°C
14. 45sec at 42°C (heat shock)
    a. One plate after the other
15. Immediately move the plate on ice for 2 minutes
16. Pour SOC medium to the reservoir
17. Transfer 80 µL of SOC medium to each well
    a. Using a multichannel pipette
    b. Discard tips after each column
18. Transfer the plate to thermoblock/PCR machine set to 37°C
19. Incubate for 1 hour shaking at 300rpm (no shaking is also working)
20. Repeat the heat shock for all PCR plates with transformed cells
21. After 1 hour of incubation, proceed with plating

Plating bacteria (~ 1 h)

1. Take the agar plates out of 4°C and let them dry (latest after the heat shock)
2. Label the plates (i.e. CL01TR_01.1a / CL01TR_01.1b & CL01TR_01.2a / CL01TR_01.2b)
    a. The square plates have 48 wells → 2 square plates for 1x 96 well plate needed
3. Place the agar plate on a paper grid with numbers and letters
    a. You will know better which grid field corresponds to which plate field
4. Add the glass beads to the grid fields (between 4-12 glass beads/ field is ok)
5. Add 70µl of the transformation to each field
    a. If you are slow it is better to work column by column
        i. Add glass beads, add bacteria, shake
        ii. You can use the lid as protection that the glass beads don't "jump" in the other column
6. Shake the plate
    a. Hold and shake the plates with both hands
    b. Check that all beads in all wells are moving
    c. Do not shake too long
7. Press the lid on the agar plate and turn the plate over
8. Take the bottom of the agar plate away
9. Transfer the glass beads in a big glass beaker
10. Clean the lid with 70% Ethanol
11. Cover the agar plate with the lid
12. Repeat steps 4-11 for all plates / columns
13. Incubate overnight at 37°C upside down
14. Add 70% ethanol to the glass beads, wash with water, transfer into a dry glass bottle and send them for autoclaving

Day 4 Colony picking and inoculation (~ 2-3 h)

Check list

- LB medium (1,5 ml per well, 150ml per plate)
- Toothpicks for picking
- Deepwell plates (Deepwell plates that are round on top and bottom, Starlab # E2896-2110)
- 1250µl digital multichannel pipette (Track) with tips

Steps:

The steps are best done with one or two additional people checking that the right well is picked and put into the correct well in the deepwell plate

1. Experimental person takes agar plates and uses computer script and enter which well has colonies (i.e. A1 - yes, A2 - no)
    a. Name of the script: script_B_picking_script.bat
    b. Can be run on lab desktop PC or via remote desktop from personal computer
    c. Takes ~ 1 hour
    d. possible break point, leave the agar plates at 4°C over the weekend
2. Use the script that makes the rearray for your experiment to create a new plate layout
    a. Name of the script:
    b. Make sure that the rearray information is saved in the expr_clone_info MySQL DB table
3. Use aseptic bench working technique
4. Label the deepwell plates (i.e. CL01GExDW_01.1a / CL01GExDW_01.1b; CL01GExDW_01.2a / CL01GExDW_01.2b)
5. Fill 1,5 ml LB-Medium in the wells
    a. Use the 1250µl digital multichannel pipette
6. Pick one colony from the first well
    a. Using a toothpick
    b. If you want to prepare 2 identical plates: stir in the corresponding well of the deep-well for a few seconds, then pick the same colony with the same toothpick into the second pick plate
    c. With the new 96 MiniPrep Kit you should get enough DNA with one deepwell plate
    d. You can leave the toothpick in the deepwell until you are done with one column
7. Continue with the next well
8. Repeat until all clones are picked
9. Cover the deepwell plate with breathable foil
10. Incubate @ 37˚C at 700rpm in the incumixer for 24h
    a. This conditions are important for successful MiniPrep

Day 5 Glycerol stock, Miniprep (~ 2 hours per plate)

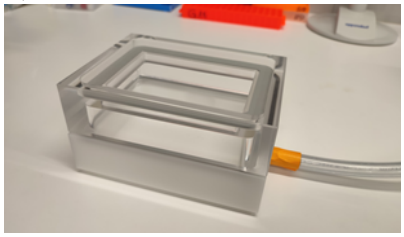
prepare glycerol stock before miniprep!

Material needed:

- 40% glycerol steril (50µl per well, 5ml per plate)
- Costar plates for glycerol stock
- Alu foil to cover glycerol stock
- 1250µl digital multichannel pipette (Track) with tips
- Qiagen 96 well Miniprep kit
- Plate inserts for big centrifuge
- Big glass beaker
- Multipette with 5ml tips
- Alu foil for resuspension
- Costar plate for elution
- Vacuum (pump set to 300 mbar)
- Waste tray = square reservoir (can be autoclaved)

Steps:

1. Work under aseptic bench working conditions
2. Get deepwell plates from the incubator
3. Prepare the glycerol stock plates (i.e. CL01GEx_01.1 & CL01GEx_01.2)
    a. By adding 50 µl of 40% glycerol to all required wells of a new costar plate
    b. Check if the bacteria are in suspension - if not, vortex (cover with alu or plastic foil before vortexing)
    c. Add 50 µl of the incubated bacteria culture to the corresponding wells and close the plate with alu-foil.
    d. Shake 30sec at 750rpm on the Thermomixer
    e. Freeze @ -80˚C
4. Centrifuge the deepwell plate @ 2100 xg for 5 min.
5. During centrifugation: Prepare the Qiavac Multiwell with Turbo filter 96 plate and S-Block QIAvac 96:



    a. Seal unused wells with additional tape
    b. Note for those using the unused well from a used plate: Because there are many vacuum steps in the procedure and the air flows better through previously-used wells (now empty) than the wells that are in use now, make sure that you tape the previously-used wells so that the airflow passes through the wells that you want. Otherwise, the air will tend to flow through the previously-used wells and reduce the efficacy of vacuum suction.
6. Pour out medium into beaker, tap dry the plate surface with paper towel
    a. If you have 2 identical deepwell plates: add the content of the second deepwell plate (CL01GExDW_01.1b) to the corresponding wells of the first plate (using digital multichannel to reduce the number of pipetting steps). Centrifuge @ 2100rpm for 5 min.
    b. Pour out the medium into a beaker
    c. Tap the plate on a paper towel to empty completely
7. Add 300 µl of buffer P1 to each well

    a. Using the multipette or digital multichannel
8. Close plate with alu foil
9. Vortex to completely resuspend the bacteria
10. Remove foil
11. Add 300 µl of buffer P2 to each well
    a. Using the multipette or digital multichannel
12. Close the plate with the plastic foil from the kit
13. Invert 6-8 times
14. Incubate 5min at room temperature.
    a. Do not let the lysis take longer than 5 min
    b. Count in time from first well having received the lysis buffer
15. Remove foil
16. Tap dry the plate top
17. Add 300 µl of buffer S3 to each well
    a. Using the multipette or multichannel
18. Close the plate with the plastic foil from the kit
19. Invert 6-8 times
20. Remove foil and tap dry the plate top
21. Transfer content of each well in the corresponding well in the Turbo Filter 96 plate
    a. Using the digital multichannel (set to 1000µl)
22. Apply vacuum
    a. Pump set to 300 mbar
    b. To suck liquid in the S-block
    c. Make sure all liquid has passed the filter plate
23. Close vacuum
24. Remove filter-plate from assembly
25. Discard the filter plate
26. Remove S-Block
    a. DNA is here
27. Install waste try in the assembly
28. Install Plasmid Plus 96 plate in the assembly
29. Seal and label unused wells with tape
30. Add 300 µl of buffer BB to each well in S-Block
    a. Using the multipette or digital multichannel
31. Close the S-Block with the plastic foil from the kit
32. Invert 1-3 times
33. Remove foil
34. Tap dry the S-Block on top
35. Transfer content of each well in the corresponding well in the Plasmid Plus 96 plate
    a. Using the digital multichannel (set to 1250µl)
36. Apply vacuum
    a. Pump set to 300 mbar
    b. To suck liquid in the waste tray
    c. Make sure all liquid has passed the plate
37. Close vacuum
38. Transfer 900 µl of buffer PE in each well in the Plasmid Plus plate
    a. Using the digital multichannel
39. Apply vacuum
    a. Pump set to 300 mbar
    b. To suck liquid in the waste tray
    c. Make sure all liquid has passed the plate
40. Close vacuum
41. Empty waste tray
42. Pat dry the nozzles of the Plasmid Plus plate until now liquid can be seen on the paper towel
43. Put back the waste tray and assemble
44. Apply vacuum for 10 min
    a. Pump set to 300 mbar
    b. To dry the filter
45. Close vacuum
46. Lift the top plate from the base - but not the Plasmid Plus plate from the top plate!

47. Vigorously tap the top plate on a stack of absorbent paper until no more drops come out
    a. Blot the nozzles of the Plasmid Plus plate with clean absorbent paper
48. Remove the waste tray
49. Place 2 “old” costar plates (one with lid and one without lid) in the assembly
    a. To reach the required height, the nozzles should reach the wells of the costar plate
50. Place your elution plate (i.e. CL01GExMP_01.1) in the assembly and reassemble
51. Add 70µl of water/EB-buffer to the center of each well of the Plasmid Plus plate
    a. Using a manual multichannel
52. Let stand for 3 min
53. Apply vacuum for 1 min
54. Close vacuum
55. Disassemble the Qiavac Multiwell to get your DNA

Stop point. DNA can be frozen @ -20˚C and stored.

Nanodrop measurement

1. Using part 1 of script C, create a template for the Nanophotometer and save it in the Nanophotometer folder on the group drive (i.e./imb-luckgr/NanoPhotometer/HTP_data/CL100/)
2. Thaw plates with DNA (i.e. CL01GExMP_01.1 & CL01GExMP_01.2)
3. Centrifuge 3min @ 3000g in the big centrifuge
4. Load the correct measuring template to the NanoPhotometer
    a. On the NanoPhotometer, click ‘Nucleic Acid’, then swipe right and click the top right button that looks like a barcode. Click ‘Sample’ and then click ‘Import’. Select ‘Network_Groupdrive’ to find the NanoPhotometer folder in the group drive mentioned in point 1. There you can find your measuring templates and load them into the Nanophotometer for measurement
5. Measure the DNA concentration
6. Save the data to the group drive in the corresponding folder
    a. Save the measurement in the same folder so that you can access it through the groupdrive too
    b. If the folder ‘Network_Groupdrive’ does not appear on the Nanophotometer, try restarting it
7. If needed you can concentrate your DNA:
    a. Place the plate (without lid) in the dessicator
    b. Turn on vacuum and let evaporate until the desired volume/concentration is reached
    c. ~36h for 20-25µl reduction in volume
    d. Ask Christian for help, .. needed

Day 6 DNA dilution and sequencing

The first sequencing is done with both backbone primers (forward and reverse), full coverage sequencing for inserts is done after results come back for those that need it

1. Use part 2 of script C to calculate the dilutions needed
2. Make sure that the measured DNA concentrations are uploaded to the expr_clone_info MySQL DB table
3. Prepare the dilutions (i.e. CL01GExDil_01.1 & CL01GExDil_01.2)
    a. according to the template you created
    b. DNA concentration should be around 100 ng/µl
4. For the expression test you will need to dilute the NL plate once more
    a. Option 1. 1:10 (you take 1µl for expression test)
    b. Option 2. 1:25 (you take 2µl for expression test)

DNA stock

1. Label PCR plates with labels for DNA stock (i.e. CL01GExSt_01.1 & CL01GExSt_01.2)
2. Pipette 10µl of the not diluted DNA (CL01GExMP_01.1 & CL01GExMP_01.2) to the stock plates
3. Close plates with alu foil and give to Mareen for storage

Sequencing

Each plate has to be submitted individually to StarSeq. You will get a zip file containing one .ab1 and .seq file for each sample submitted in the plate. You can use the plate barcodes for plates with more than 78 samples or you can submit individual barcodes for plates with <78 samples. If you are sending a plate to Starseq you have to have at least 48 samples on the plate

Steps:

1. Prepare an Excel file (one for each sequencing run) with the file names of your sequencing samples in 96-well format. Suggested file names: e.g. mCit-[ORF ID]-F for the mCit construct and forward read. The layout should correspond to what you have generated after picking the colonies (i.e. CL01GExh_01.1)
2. Label the PCR plates for sequencing (i.e. CL01GExSF_01.1 & CL01GExSF_01.2; CL01GExSR_01.1 & CL01GExSR_01.2).
3. Add 1µl of the corresponding primer to the plates
    a. primer # 44 NanoLuc-398fwd - for N-terminal NL fusion
    b. primer # 47 mCitrine-547fwd for N-terminal mCit fusion
    c. primer # 51 pEXP_rev for no C-terminal fusion
    d. Using the multipette and combitip 1ml
    e. Alternatively, one can also aliquot the primers into PCR tubes and use digital multichannel to distribute the primers into the wells
4. Add 6 µl of the diluted DNA to the sequencing plate
    a. I.e. from CL01GExDil_01.1 & CL01GExDil_01.2
    b. Using manual multichannel pipette
5. Close the sequencing plate using the alu foil
6. Order the sequencing on the StarSeq webpage
    a. Use the Excel file created in step 1 to copy paste the plate layout into their web form
7. Pack plate together with paperwork in a padded envelope
    a. To avoid the foil getting pierced
    b. Submit each plate as an individual sequencing run
    c. When submitting multiple plates, results will likely not come back all by next morning but over the next 24-36h
8. Process the sequencing results with the Sanger seq processing pipeline
    a. Instructions can be found in labfolder under templates
9. Make sure to update results accordingly in the expr_clone_info MySQL DB table

Day 7 Transfection

Expression test

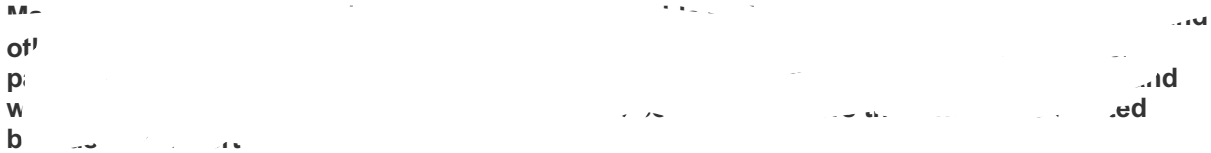
CS notes:

- I did get $6x10^6$ HEK293 cells out of 1 T-25 flask lately.
- I found it more convenient to do the triplicates in separate plates.
- I did not mix DNA with Lipofectamin before, only when I put the DNA to the final incubation plate.
- While I did NL and mCit the same day, I pipetted them separately as it is very hard to handle 6 plates at the same time.
- The volumes I put here (most of the time) depend on your transfection ratio and DNA concentrations used. I did NL-constructs 4ng/µl, mCit 100ng/µl, pcDNA 200ng/µl; 2:50 ratio

Steps:

1. Prepare the layout of your plate with the controls
    a. controls: NL-stop + pcDNA, mCit-stop + pcDNA, well with only pcDNA, well with only cells
    b. The controls you put depend on your experiment and space you have on the plate. I[illegible] you ha[illegible] to Katja
2. Prepare the DNA for your controls, PA-mCit-Stop, NL-Stop and pcDNA3.1
    a. can be prepared in PCR stripes - then you can later use the multichannel pipette
3. Prepare an additional dilution of the NL-constructs to 4ng/µl (if you haven't already)
4. Take a PCR plate
5. Add the pcDNA (3µl) to the wells first.
    a. Using multipette or multichannel pipette
    b. Doing the pcDNA first allows you to do everything with one tip. Try to get the DNA to the bottom of the plate.
6. Add the **NL-Stop** (for the mCit-constructs) **or mCit-Stop** (for the NL-constructs)
    a. 2µl to the wells
    b. Using the multipette or multichannel pipette
    c. for multipette: using one tip is possible for this as the only possible contamination would be with pcDNA which can be avoided by putting the DNA at the wall of the wells away from the pcDNA
7. Add your diluted construct DNA (mCit or NL)
    a. 2µl if you are using the DNA concentrations written on top
    b. Using the multichannel pipette
8. Add the DNA for the controls to the wells
9. Tap plate to mix all DNA in the bottom of the well
10. Add 100µl Optimem to each well
11. Prepare the Lipofectamine-Optimem mixture in a 15ml Falcon tube
    a. You do not need to do quadruples here. This saves some lipofectamine
    b. Example:
    78 wells/plate x 0.5µl Lipo/well x 3 plates = 117µl Lipo
    78 wells/plate x 25µl Optimem/well x 3 plates = 5.85ml Optimem
    Now add some for the reservoir:
    => 120µl Lipofectamine + 6ml Optimem
12. Label the plates for incubation (i.e. LuXXXrXX)
13. Add Lipo-Opti mixture to a 10ml reservoir by pipetting
    a. Decanting is suboptimal as it leaves some residual mixture in the Falcon tube
14. Add 25µl Lipo-Opti mixture to each well of the incubation plates (white 96well plate for LuTHy)
    a. Using the multichannel pipette
15. Take out cells, wash and add trypsin
16. While the trypsination is ongoing:
    a. Transfer DNA-Opti mixture in the incubation plates
    b. You can use a digital multichannel (Trick) to speed it up (aspirate 75µl, dispense 3x25µl)
    c. Predispense step is needed to get accurate amounts for the first dispense of the multi-dispense

    d. The 20min time limit starts now
17. Quench trypsin, resuspend cells, count cells, centrifuge and adjust concentration to $2.67x10^5$ cells/ml in phenol-red free DMEM medium.
18. Decant the cells in a 25ml reservoir
19. Add 150µl cell suspension to the plates
    a. Using the digital multichannel
    b. aspire 450µl, then dispense 3x150µl doing the triplicate without changing tips
    c. Use program called" LUTHY CELLS" in 1250µl digital multichannel pipette (Track). The program first resuspends the cell multiple times (called 'Mix' in the program), and then aspirates 450µl for the repeat dispense of 3x150µl
20. Incubate for 48h
21. Proceed with measurement as usual for LuThy assay
22. For the LuTHy processing scripts to be able to process your data, KL numbers have to be generated for all the constructs on your plate.
23. Make sure the KL numbers are generated and saved in the LUCK_DB.Luck_lab_plasmids table along with all available information.



5.1.2 The medium-throughput site-directed mutagenesis

Site-directed mutagenesis (without Kit)

Day 0 Primer design

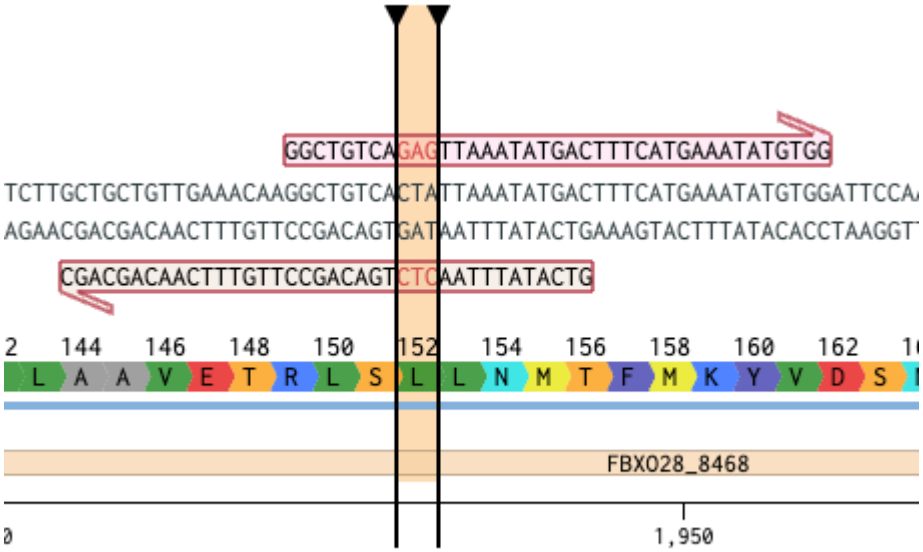
Criteria for mutagenesis primers:

- Primer length should be 32-36 nt. If it is shorter, the mutation might not be cloned properly!
- GC content of primer should be between 40 to 60%
- Difference in melting temperature between the forward and reverse primers should be ideally less than 5ºC. (use NEB Tm calculator: https://tmcalculator.neb.com/#!/main and select Phusion as the product group for the melting temperature of primer)
- The annealing temperature of PCR reaction should be set at the value which corresponds to 5ºC lower than the lowest melting temperature among the primers
- The 3` end of the primer should ideally be C or G
- The annealing temperature should be below 70°C if possible

primer order info, if you have 24 or more primers (IDT company)
If you want to orde primers in plate
price wise:
The prices for DNA oligos when in tubes or in plates can be seen in our website here . The prices are usually a bit lower for oligos in plates however, one should look at the final cost of the whole order. For example when ordering in plates there is a minimum of 24 oligos that should occupy the plate. So in the long run, plates are not always the cheaper option.
dry or wet primers
When ordering in plates you can choose your oligos to be normalized to a certain amount. That can be either as a pellet (dry) or resuspended (wet). In that way you will avoid needing higher volume than the capacity of the well. These can be adjusted in the "Plate Specifications" button when ordering the plate oligo . I would say that there are no pros and cons for primers in pellet or in solution in terms of primer performance, stability and so on. It is more a matter of experimental needs and set up. Some researchers prefer to receive their oligos ready to use whereas others want to resuspend them in a certain buffer or in a specific dilution. When automation and robot handling is included, people prefer having plates than tubes. On the other hand, when having the oligos in plates and manual pipetting is done the chances for contamination or spillage can be higher.

Design of a point mutation (non-kit way)

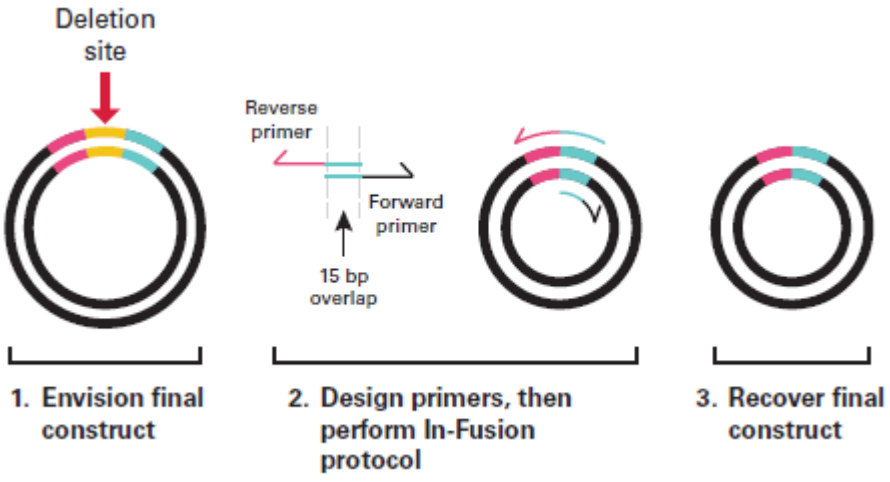
- Design forward and reverse primers that overlap at the site of mutation. Try to locate the mutation to be at the middle of the overlapping region so that the mutation is flanked by complementary sequences. The overlapping part (that contains your mutation) should be 20-2215-22nt nt long.
- Here is an example to mutate L152E in benchling. To know what codon codes for E, right click on L152 and select ‘Change amino acid’. Remember to change Organism to Homo sapiens. There you can find the codon that codes for the amino acid that you want to mutate to, and the best codon change to achieve that amino acid substitution. Do take into consideration the number of bases that need to be changed for the amino acid substitution and the frequency of codon to ensure optimal mutagenesis



-

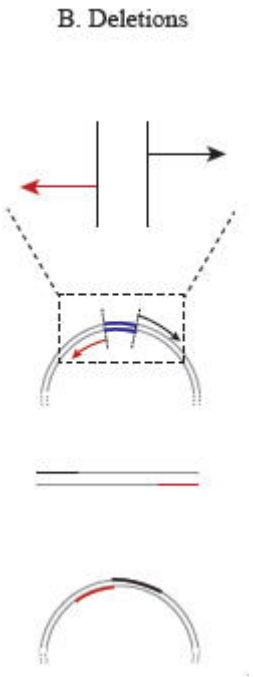
Design a deletion (non kit way)

- The forward and reverse primers have to overlap at the overhang so that the synthesized strands can circularize after amplification. The non-overhang region should have ~20 nt and the overhang (overlap) region should have 15 nt.
- Here is a schematic showing how the primers with overhang should be designed. (the scheme and explanation are retrieved from takara : https://www.takarabio.com/learning-centers/cloning/applications-and-technical-notes/mutagenesis-with-in-fusion-cloning)



Design a Deletion with Q5 kit

- Design forward and reverse primers that exclude the deletion site.
- Here is the schematic.



Design sequencing primer

- use primer design tool
- follow the instruction of the primer designer tool
    - in labfolder → “templates” → “instructions” → “How to design primers with PrimerDesigner”

Primer order from Sigma

1. design the primer
2. order Oligos in solution (water)
    a. the price is the same and it will save time to you
3. import all needed data into the excel file from sigma
    a. there are 2 excel files: 1x for single oligo order and 1x to order oligos in plate
    b. can be found on the group drive (Primer_AG_Luck → Sigma_order_template or DNAPLATE_96well_8ch_template_sigma)
    c. the excel sheet to order single oligos is also saved on the intranet (Administration → Purchasing → Oligo Ordering → Sigma Oligo template)
    d. the link to order oligos in plate: [DNA-Oligos in Platten (sigmaaldrich.com)](https://www.sigmaaldrich.com/DE/de/configurators/plate?product=dnaplate&activeLink=sequenceUpload)
    e. Oligos in plate:
        i. ask for quote for your order
        ii. after uploading the file you need to set the “scale”. “purification” and “format”
        iii. scale = 0,025; purification = desalt; format = in solution (water)
4. upload the excel sheet to the order manager under service agreements
    a. oligos are ordered every tuesday and thursday after 2pm
    b. it might take up to 1 week to receive the oligos, oligos in plate take ~ 2days longer than oligos in tubes
5. after receiving the oligos you have to dilute them

Primer order from IDT

1. design the primer
2. register with IDT
3. go to “Products and Service” → “DNA and “RNA” → “Custom DNA oligos” → “DNA Oligos” → “Single-stranded DNA” (you can choose between tubes and plates) → press the button “order now” →
4. order primer in tubes:
    a. enter all required informations in the fields

    b. you can use the button “bulk input” if you have several oligos to order
        i. Scale: 25nmole DNA oligo
        ii. Formulation: you can choose “None” = dry or “LabReady (100µM in IDTE, pH8,0)”
        iii. Purification: Standard desalting
        iv. Sequence:
5. order primer in plates:
    a. choose plates
    b. 25nmole → order
    c. download the excel sample ordering template (under upload plates)
    d. fill out the excel sheet
    e. upload the excel sheet
    f. check the upload
    g. if necessary make changes
6. add to order
7. order the oligos via internet, add the email of the purchase department “**einkauf@imb-mainz.de”** in the distribution list for order confirmations
8. enter your IDT order in the Order Manager

Primer dilution in plates

1. take a new PCR plate
2. label the plate (i.e. MU01PrDilF_01 and MU01PrDilR_01)
3. add 90µl water in each well
4. add 10µl primer in the corresponding well
5. close the plate
6. mix/vortex
7. freeze until needed

Preparation of template DNA 10ng/µl (i.e. MU01TD_01)

1. take a new PCR plate
2. label the plate (i.e. MU01TD_01)
3. add 9µl water in each well
4. add 1µl template DNA in the corresponding well
    a. take the template DNA from your diluted MiniPrep 100ng/µl
5. close the plate
6. mix/vortex
7. freeze until needed

Day 1 PCR, DPN1 digestion and E-Gel

Checklist PCR, DPN1 digestion and E-Gel:

- ☐ 96-well skirted PCR plates (3x)
- ☐ 5ml tube (Axygen, # SCT-5ml-S) or 50 mL falcon tube - to prepare PCR master mix, 50ml because of multipette
- ☐ PCR foil
- ☐ multichannel pipette 10µl
- ☐ 10µl pipette tips (4 boxes)
- ☐ multichannel pipette 50µl
- ☐ 100µl tips
- ☐ multipette
- ☐ combitips 1ml (to add the PCR Master Mix)
- ☐ combitips 0,1ml (to add DPN1)
- ☐ 100µl pipette
- ☐ 100µl pipette tips
- ☐ 1000µl pipette
- ☐ 1000µl pipette tips
- ☐ ice block - to keep PCR components in cold

- ☐ PCR machine or Thermomixer
- ☐ PCR components
- ☐ DPN1
- ☐ E-Gel 96 1% Agarose (GP) (invitrogen, # G700801)
- ☐ E-Gel 96 High range DNA marker

PCR program:

temperature	time	cycle	step
98˚C	2min	1x	initial denaturation
98˚C	30s	25x	denaturation
__ __˚C *	15s	25x	primer annealing
72˚C	5min (1min/1kb)	25x	extension
72˚C	5min	1x	final extension
16°C	∞	1x	

* Temperature depends on the primer, try to keep the temperature below 70°C when designing the primer
if melting temp of 1primer is less than 69°C than annealing temp = 55°Cif higher = 63°C

PCR reaction (50µl total)

PCR components	1x (1well)	x 100
primer (10µM)	2.5 µL	250 µL
primer (10µM)	2.5 µL	250 µL
template DNA (10ng)	1µl	100 µl
dNTPs (10mM)	1 µL	100 µL
10x HF Buffer	10 µL	1000 µL
High fidelity DNA polymerase	0.5 µL	55 µL
H_2O	32.5 µL	3250 µL (= 5x 650 µL)

Master Mix

PCR components	1x (1well)	x 100
dNTPs (10mM)	1 µL	100 µL
10x HF Buffer	10 µL	1000 µL
High fidelity DNA polymerase	0.55 µL	55 µL
H_2O	32.5 µL	3250 µL (= 5x 650 µL)

Steps:

1. Label the PCR plate (i.e. MU01PCR_01)
2. Once the PCR components started to thaw vortex each PCR reagent
3. Prepare the master mix
    a. In 5ml tube or 50mL falcon tube
4. Pipette 44 µl of the master mix in each well of the PCR plate (on ice/cold block)
    a. 44,5µl is not possible with multipette
    b. Using the multipette and combitip 1ml
5. Add 2,5µl of each primer to the PCR plate
    a. Using the multichannel pipette
    b. Pipette from the primer working solution plate (i.e. MU01PrDilF_01, MU01PrDilR_01)
6. Add 1µl of purified template DNA (~10ng)
    a. Use multichannel pipette
    b. Pipette from the template DNA plate (i.e. MU01TD_01)
7. One well should be used as control (master mix without ORF)
8. Close the PCR plate with PCR foil
    a. be sure to close every column and row using the grey plastic “card”
9. Vortex the plate briefly
10. Centrifuge briefly
11. Run the PCR (~3 hours)

if melting temp of 1primer is less than 69°C than annealing temp = 55°C

if higher = 63°C

DpnI digestion (using commercial DPN1)

Steps:

1. Prepare a new PCR plate with DPN1
    a. Can be done while the PCR is running
    b. Label the plate (i.e. MU01Dpn_01)
    c. Add 2µl DPN1 with the multipette to the DPN1 plate
        i. The multipette **must** touch the PCR plate while pipetting to ensure that the 2µl of DPN1 enters into each well
2. Add 50µl of PCR product to the plate with DPN1
    a. Using the multichannel pipette 50µl
3. Incubate for 1h at 37°C (PCR machine or thermomixer)
4. Incubate for 20 min at 65°C (to stop the DPN1 reaction)

Validation of the PCR product with E-gel

- Info:
    - PCR products can be stored at 4°C for 48h, for longer time freeze PCR products
    - Document all wells that do not look ok on gel -

Steps:

1. Label the E-Gel plate (i.e. MU01Gel_01)
2. Pipette 25 µl of blue 96 gel loading buffer in the E-Gel plate
    a. Using the multipette and 2,5ml Combitip
    b. Can be done while PCR is running
3. Add 6 µl of PCR product to each well
    a. Using the 10µl multichannel pipette
4. Install 96 well E-gel to the motherbase

5. Load 20µl PCR/buffer mix to each well
    a. Using the 50µl multichannel pipette
6. Load 20µl of E-Gel 96 High range DNA marker
7. All empty wells must also be filled with 20µl
    a. With buffer or loading dye
8. Insert the plug into the socket
9. Run gel for 12 min
    a. Program EG
10. Take picture with GelDoc Station
11. Analyze gel picture with the E-Editor 2.0 software
    a. On the desktop PC in the technical room
    b. Realign the bands and save it in your cloning project folder
    c. The software is pretty self-explanatory and has a manual available under the help button. Ask Katja for help.
    d. explanation how to do it by john
12. Decide if PCR was successful and whether it is worth proceeding
13. Document all wells that did not look ok
    a. Add this information to the respective MySQL table

Preparing square agar plate (should be done at least the day before needed)

Check list

- ☐ LB Agar (250ml/square plate)
- ☐ Square plates and divider
- ☐ Microwave

1. Take LB-Agar (250ml) from IMB media lab
2. Use aseptic bench working technique
3. Heat Agar in the microwave (program: soften/melt, 2= melt dark chocolate, 100 = 5,5 min; after 3x the agar is liquid)
4. Let it cool down (i.e. add a clean stirrer to the agar and place the bottle on the magnetic stirrer, adjust the temperature to 50°C and 250rpm)
5. Add antibiotic (250µl) when the agar is cooled down sufficiently and you are ready to pour the plates
6. Take out the plate from the plastic protection
7. Add agar to the plate (pop bubbles with a pipette tip or move them to the side)
8. Take out the grid from the plastic protection
9. Add the grid in the square plate with agar
   --> the grid does not stay down - weigh down the grid with something (i.e. a 250ml bottle)
10. Let the agar solidify
11. Store at 4°C (upside down)

Day 2 Transformation and plating

Checklist Transformation and plating:

- ☐ 48 well square plates with agar and antibiotic (2 plates are needed for 96 well plate)
- ☐ SOC medium (8ml/plate)
- ☐ 10ml reservoir
- ☐ DH5a (30µl)
- ☐ multichannel pipette 50µl
- ☐ 100µl pipette tips
- ☐ multichannel pipette 300µl
- ☐ 300µl pipette tips
- ☐ 200µl pipette
- ☐ 200µl pipette tips

- [ ] glass beads
- [ ] 70% Ethanol
- [ ] Thermomixer/PCR machine at 42°C and 37°C
- [ ] Ice box

Transformation:

1. Take out SOC medium (for one well = 80 µL, for 1 plate = 8 mL) and let it thaw at room temperature
    a. 50ml takes long time to thaw, could be placed at 4°C the afternoon before
2. Use aseptic bench working technique
3. Take out DH5α from -80°C
    a. Put them immediately on the ice
    b. Let them thaw
    c. Label the plate (i.e. MU01_TR01)
4. Take the plate after Dpn1 digestion (i.e. MU01Dpn_01)
5. Transfer the plates on ice
6. Transfer of the digested PCR product into the DH5a plate
    a. Us3 µL e a multichannel pipette
    b. No resuspension, no vortex when adding the PCR product into the DH5a
    c. Close the plate with alu foil
7. Incubate for 30 minutes on ice (bacteria with PCR product)
8. **Meanwhile**: set the thermoblock to 42°C for the heat shock and set another thermoblock to 37°C
9. 45sec at 42°C (heat shock)
10. Immediately move the plate on ice for 2 minutes
11. Pour SOC medium to the reservoir
12. Transfer 80 µL of SOC medium to each well
    a. Using a multichannel pipette
    b. Discard tips after each column
13. Transfer the plate to thermoblock to 37°C
14. Incubate for 1 hour shaking at 300 rpm (no shaking is also working)
15. After 1 hour of incubation, proceed with plating

Plating bacteria (~ 1 h)

1. Take the agar plates out of 4°C and let them dry (latest after the heat shock)
2. Label the plates (i.e. MU01_TR_01a, MU01TR_01b)
    a. The square plates have 48 wells → 2 square plates for 1x 96 well plate needed
3. Place the agar plate on a paper grid with numbers and letters
    a. You will know better which grid field corresponds to which plate field
4. Add the glass beads to the grid fields (between 4-12 glass beads/ field is ok)
5. Add 70µl of the transformation to each field
    a. 70µl needs a bit longer to dry - do not turn immediately after shaking
    b. If you are slow it is better to work column by column
        i. Add glass beads, add bacteria, shake
        ii. You can use the lid as protection that the glass beads don't "jump" in the other column
6. Shake the plate
    a. Hold and shake the plates with both hands
    b. Check that all beads in all wells are moving
    c. Do not shake too long
7. Press the lid on the agar plate and turn the plate over
8. Take the bottom of the agar plate away
9. Transfer the glass beads in a big glass beaker
10. Clean the lid with 70% Ethanol
11. Cover the agar plate with the lid
12. Repeat steps 4-11 for all plates / columns

13. Incubate overnight at 37°C upside down
14. Add 70% ethanol to the glass beads, wash with water, transfer into a dry glass bottle and send them for autoclaving

Day 3 Colony picking and inoculation (~ 2-3 h)

Check list

- ☐ LB medium (1,5 ml per well, 150ml per plate)
- ☐ Toothpicks for picking
- ☐ Deepwell plates (Deepwell plates that are round on top and bottom, Starlab # E2896-2110)
- ☐ 1250µl digital multichannel pipette (Track) with tips

Steps:

The steps are best done with one or two additional people checking that the right well is picked and put into the correct well in the deepwell plate

1. Experimental person takes agar plates and uses computer script and enter which well has colonies (i.e. A1 - yes, A2 - no)
    a. Name of the script: script_B_picking_script.bat
    b. Can be run on lab desktop PC or via remote desktop from personal computer
    c. Takes ~ 1 hour
    d. possible break point, leave the agar plates at 4°C over the weekend
2. Use the script that makes the rearray for your experiment to create a new plate layout
    a. Name of the script:
    b. Make sure that the rearray information is saved in respective MySQL table
3. Use aseptic bench working technique
4. Label the deepwell plates (i.e.MU01DW_01)
5. Fill 1,5 ml LB-Medium in the wells
    a. Use the 1250µl digital multichannel pipette
6. Pick one colony from the first well
    a. Using a toothpick
    b. If you want to prepare 2 identical plates: stir in the corresponding well of the deep-well for a few seconds, then pick the same colony with the same toothpick into the second pick plate
    c. With the new 96 MiniPrep Kit you should get enough DNA with one deepwell plate
    d. You can leave the toothpick in the deepwell until you are done with one column
7. Continue with the next well
8. Repeat until all clones are picked
9. Cover the deepwell plate with breathable foil
10. Incubate @ 37˚C at 700rpm in the incumixer for 24h
    a. This conditions are important for successful MiniPrep

Day 4 96 well Miniprep and Nanodrop measurement

for the MiniPrep please use the protocol “Miniprep_96well_plate” in labfolder

Day 5 DNA dilution and sanger sequencing

5.1.3 Figures

Expression profiles

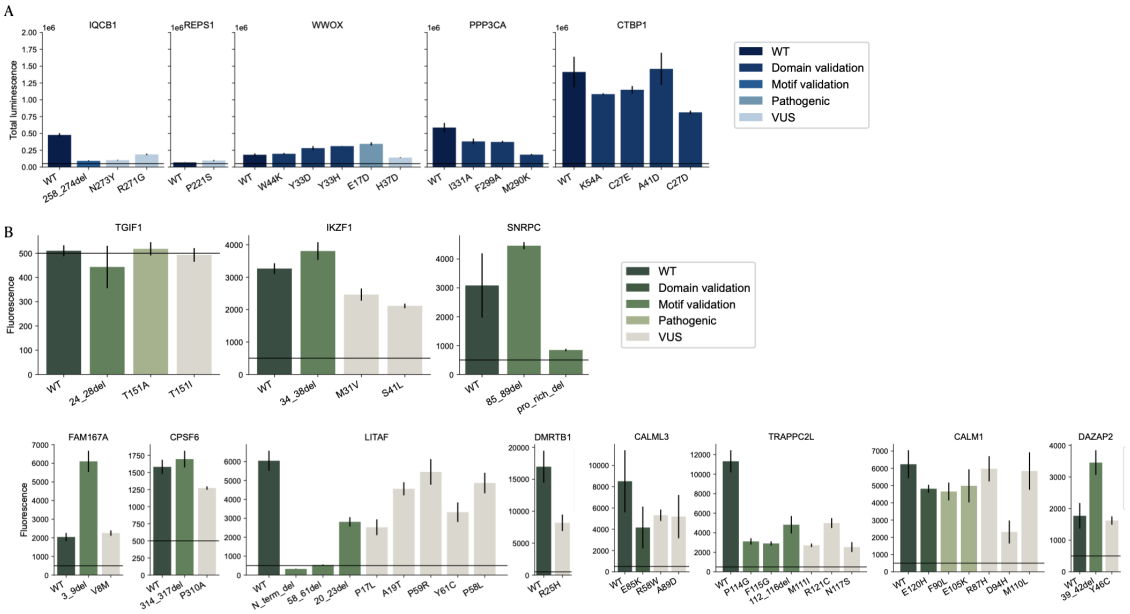


Figure 5.1: Expression of wild-type proteins and mutants. (A) The expression levels of wild-type (WT), mutants, and patient variants fused to NanoLuc were measured. The x-axis represents the names of the wild-type, mutants, and variants, while the y-axis indicates the luminescence intensity for each protein. Each protein was co-expressed with an empty mCit-control to verify expression. (B) The expression levels of wild-type (WT), mutants, and patient variants fused to mCit were assessed. The x-axis represents the names of the wild-type, mutants, and variants, while the y-axis shows the fluorescence intensity for each protein. To verify expression, each protein was co-expressed with an empty NL control.

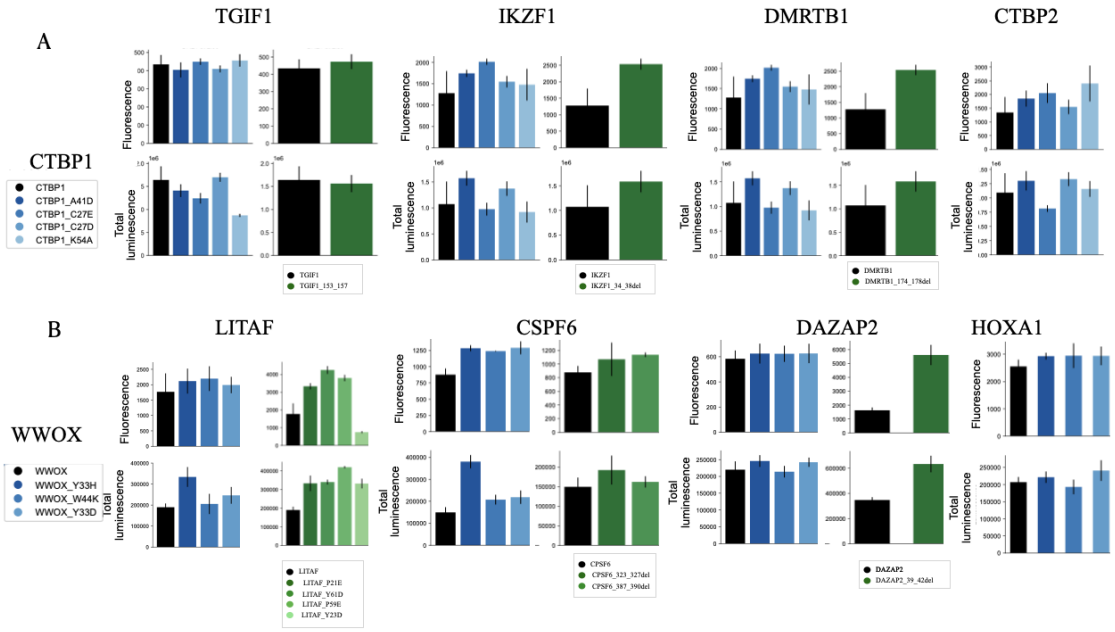


Figure 5.2: Expression of wild-type proteins and mutants protein pairs during BRET saturation assay. (A) The bar plots indicate the luminescence intensity for NL-CTBP1 wild-type and mutants, and the fluorescence intensity for mCit fused partners wild-type and mutants as well. (B) The bar plots indicate the luminescence intensity for NL-WWOX wild-type and mutants, and the fluorescence intensity for mCit fused partners wild-type and mutants.

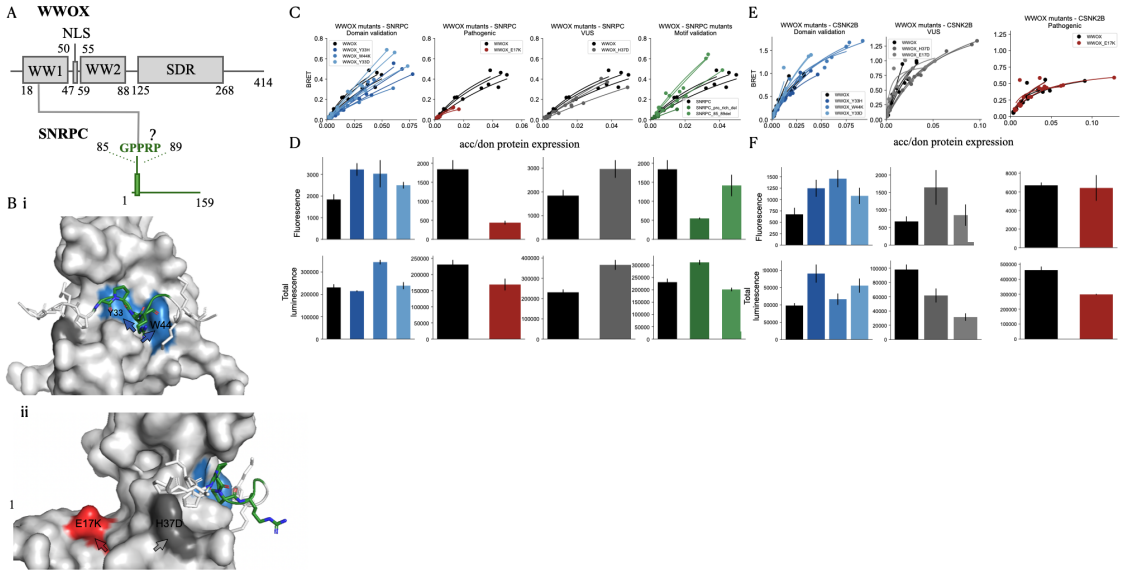


Figure 5.3: The validation of predicted interface of WWOX-SNRPC interaction and the variant effect on this ppi(A) Schematic representation of the WWOX-SNRPC interaction and putative interface. The protein containing the predicted interacting motif is shown in green, and the domain-containing protein is in grey. The question mark indicates that the AF-MM fragmentation approach was used to predict the potential interface. B) AF-MM predicted interface structural models: (Bi) The WWOX WW domain (grey) with highlighted mutated residues (blue) for domain validation and the motif (green). (Bii) The same predicted structure illustrating pathogenic (red) and VUS (grey) variants in the WWOX WW domain. (C-D) BRET saturation assay data and expression profiles: (C) BRET saturation curves showing the effects of WW domain mutants (see legend), motif deletion, and N-terminal truncation of SNRPC on binding affinity. The effects of pathogenic (red) and VUS (grey) variants on the interaction are also shown. (D) Expression profiles of wild-type and mutant interactions, with color coding corresponding to panel (C). (E-F) Validation of the WWOX-CSNK2B interaction:(E) BRET saturation curves showing the effects of WW domain mutants (see legend) and pathogenic (red) and VUS (grey) variants on the interaction with CSNK2B. (F) Expression profiles of wild-type and mutant interactions in the WWOX-CSNK2B interaction, color-coded as in panel (E).

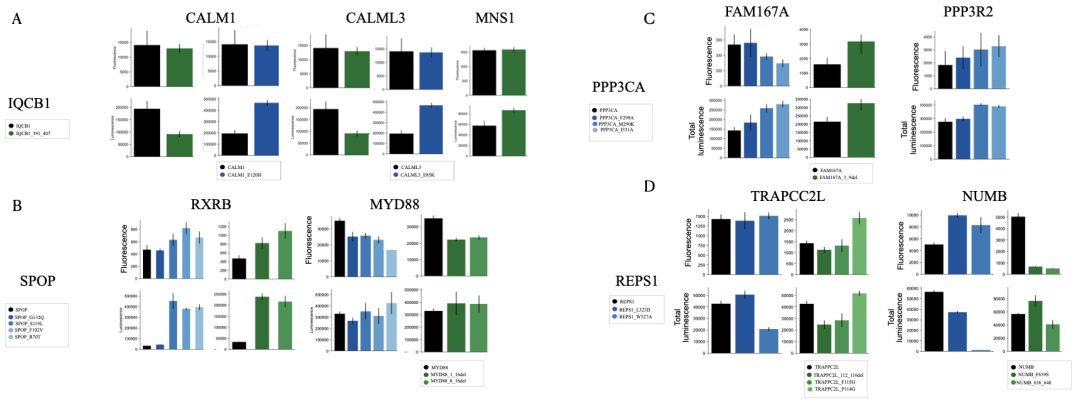


Figure 5.4: Expression of wild-type proteins and mutants protein pairs during BRET saturation assay. (A) The bar plots indicate the luminescence intensity for NL-IQCB1 wild-type and mutants, and the fluorescence intensity for mCit fused partners wild-type and mutants as well. (B) The bar plots indicate the luminescence intensity for NL-SPOP wild-type and mutants, and the fluorescence intensity for mCit fused partners wild-type and mutants. (C) The bar plots indicate the luminescence intensity for NL-PPP3CA wild-type and mutants, and the fluorescence intensity for mCit fused partners wild-type and mutants. (D) The bar plots indicate the luminescence intensity for NL-REPS1 wild-type and mutants, and the fluorescence intensity for mCit fused partners wild-type and mutants.

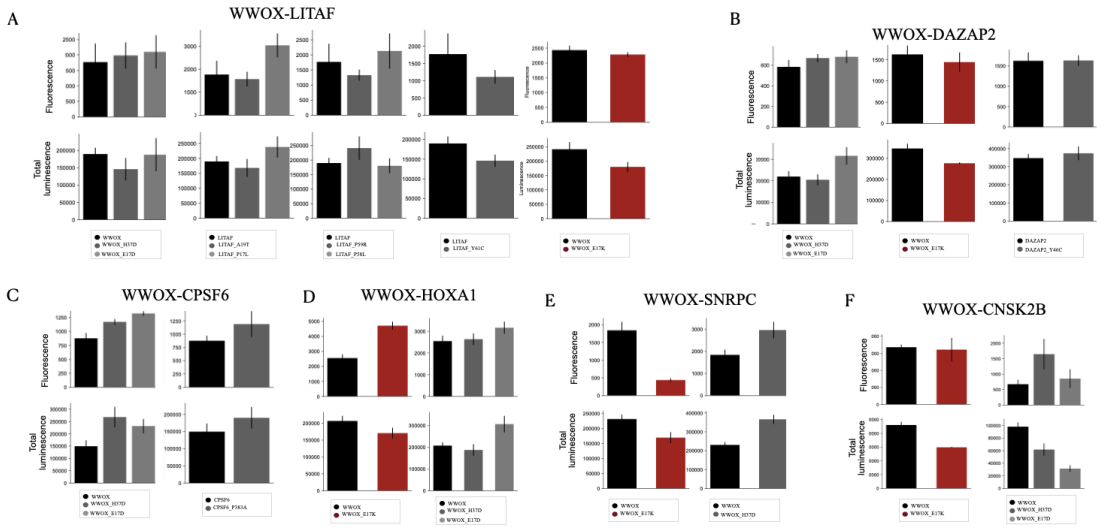


Figure 5.5: Expression of wild-type proteins and variants pairs during BRET saturation assay. (A)The bar plots indicate the luminescence intensity for NL-WWOX wild-type and variants, and the fluorescence intensity for mCit-LITAF wild-type and variants as well. (B) The bar plots show the expression levels of WWOX-DAZAP2 wild-type and variant interactions. The luminescence intensity for NL-WWOX wild-type and mutants, and the fluorescence intensity for mCit-DAZAP2 wildtype and VUS Y46C. (C) The bar plots indicate the luminescence intensity for NL-WWOX wild-type and variants, and the fluorescence intensity for mCit-CPSF6 wild-type and VUS. (D) The bar plots indicate the luminescence intensity for NL-WWOX wild-type and variants, and the fluorescence intensity for mCit-HOXA1 wild-type. (E)The bar plots indicate the luminescence intensity for NL-WWOX wild-type and variants, and the fluorescence intensity for mCit-SNRPC wild-type. (F) The bar plots indicate the luminescence intensity for NL-WWOX wild-type and variants, and the fluorescence intensity for mCit-CNSK2B wild-type.

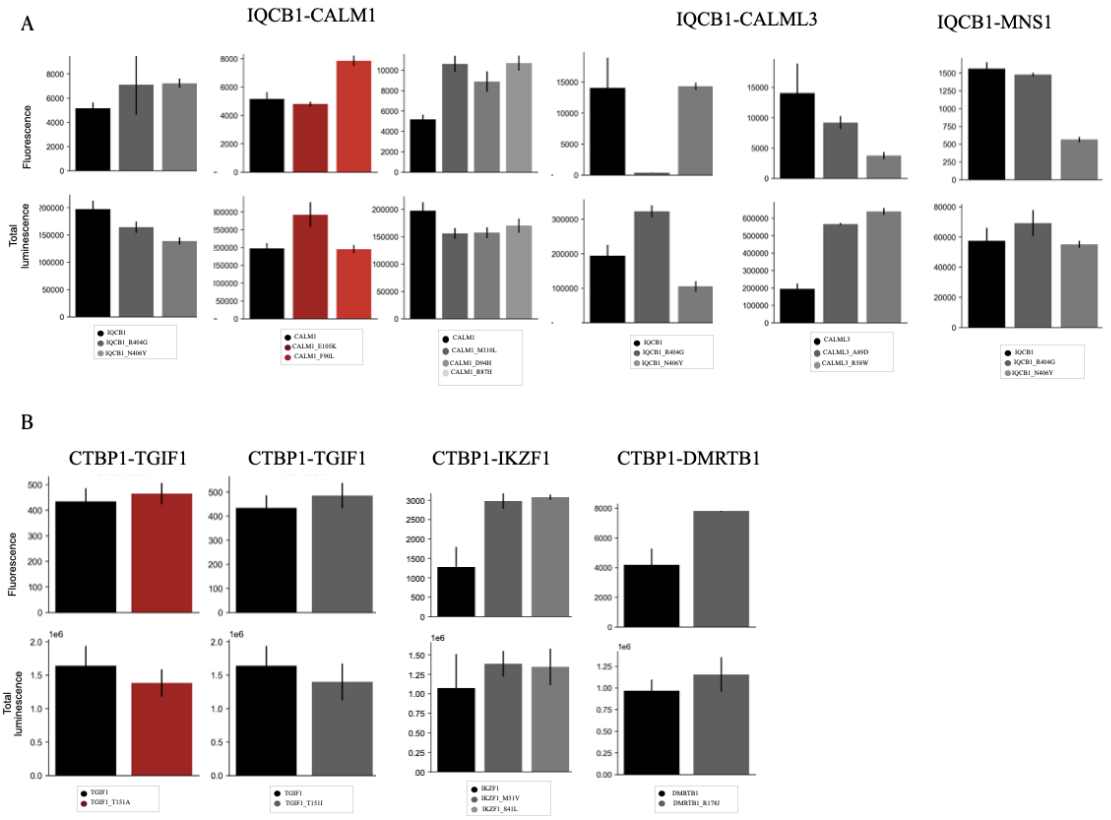


Figure 5.6: Expression of wild-type proteins and variants pairs during BRET saturation assay. (A)The bar plots indicate the luminescence intensity for NL-IQCB1 wild-type and variants, and the fluorescence intensity for mCit-fused partners wild-type and variants. (B) The bar plots indicate the luminescence intensity for NL-CTBP1 wild-type and variants, and the fluorescence intensity for mCit-fused partners wild-type and variants.

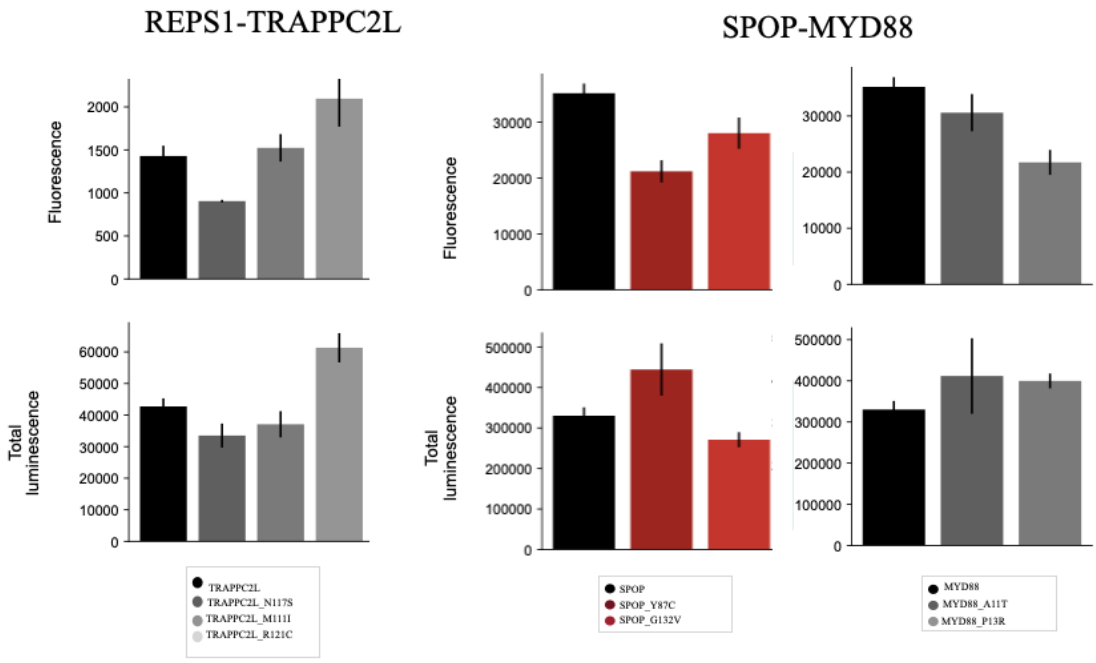


Figure 5.7: Expression of wild-type proteins and variants pairs during BRET saturation assay. The bar plots indicate the luminescence intensity for NL-REPS1 wild-type and variants of TRAPPC2L, the luminescence intensity for NL-SPOP wild-type and variants, and the fluorescence intensity for mCit-MYD88 partners wild-type and variants.

Bibliography

Adzhubei, Ivan A, Steffen Schmidt, Leonid Peshkin, Vasily E Ramensky, Anna Gerasimova, Peer Bork, Alexey S Kondrashov, and Shamil R Sunyaev (2010). "A method and server for predicting damaging missense mutations." In: *Nature Methods* 7.4, pp. 248–249. DOI: 10.1038/nmeth0410-248.

Akdel, Mehmet, Douglas E V Pires, Eduard Porta Pardo, Jürgen Jänes, Arthur O Zalevsky, Bálint Mészáros, Patrick Bryant, Lydia L Good, Roman A Laskowski, Gabriele Pozzati, Aditi Shenoy, Wensi Zhu, Petras Kundrotas, Victoria Ruiz Serra, Carlos H M Rodrigues, Alistair S Dunham, David Burke, Neera Borkakoti, Sameer Velankar, Adam Frost, Jérôme Basquin, Kresten Lindorff-Larsen, Alex Bateman, Andrey V Kajava, Alfonso Valencia, Sergey Ovchinnikov, Janani Durairaj, David B Ascher, Janet M Thornton, Norman E Davey, Amelie Stein, Arne Elofsson, Tristan I Croll, and Pedro Beltrao (2022). "A structural biology community assessment of AlphaFold2 applications." In: *Nature Structural & Molecular Biology* 29.11, pp. 1056–1067. ISSN: 1545-9993. DOI: 10.1038/s41594-022-00849-w.

Alberts, Bruce, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter (2002). *Molecular Biology of the Cell*. Garland Science. ISBN: 0-8153-3218-1, 0-8153-4072-9.

Apic, G, J Gough, and S A Teichmann (2001). "Domain combinations in archaeal, eubacterial and eukaryotic proteomes." In: *Journal of Molecular Biology* 310.2, pp. 311–325. DOI: 10.1006/jmbi.2001.4776.

Arimura, T, T Nakamura, S Hiroi, M Satoh, M Takahashi, N Ohbuchi, K Ueda, T Nouchi, N Yamaguchi, J Akai, A Matsumori, S Sasayama, and A Kimura (2000). "Characterization of the human nebulette gene: a polymorphism in an actin-binding motif is associated with nonfamilial idiopathic dilated cardiomyopathy." In: *Human Genetics* 107.5, pp. 440–451. DOI: 10.1007/s004390000389.

Babu, M Madan, Richard W Kriwacki, and Rohit V Pappu (2012). "Structural biology. Versatility from protein disorder." In: *Science* 337.6101, pp. 1460–1461. DOI: 10.1126/science.1228775.

Babu, M Madan, Robin van der Lee, Natalia Sanchez de Groot, and Jörg Gsponer (2011). "Intrinsically disordered proteins: regulation and disease." In: *Current*

Opinion in Structural Biology 21.3, pp. 432–440. DOI: 10.1016/j.sbi.2011.03.011.

Bagowski, Christoph P., Wouter Bruins, and Aartjan J. W. Te Velthuis (2010). "The nature of protein domain evolution: shaping the interaction network". In: *Current Genomics* 11.5, pp. 368–376. DOI: 10.2174/138920210791616725.

Berg, J M and H A Godwin (1997). "Lessons from zinc-binding peptides." In: *Annual review of biophysics and biomolecular structure* 26, pp. 357–371. DOI: 10.1146/annurev.biophys.26.1.357.

Björklund, Asa K, Diana Ekman, Sara Light, Johannes Frey-Skött, and Arne Elofsson (2005). "Domain rearrangements in protein evolution." In: *Journal of Molecular Biology* 353.4, pp. 911–923. DOI: 10.1016/j.jmb.2005.08.067.

Blake, C. C. F., D. F. Koenig, G. A. Mair, A. C. T. North, D. C. Phillips, and V. R. Sarma (1965). "Structure of Hen Egg-White Lysozyme: A Three-dimensional Fourier Synthesis at 2 Å Resolution". In: *Nature* 206, pp. 757–761. DOI: 10.1038/206757a0.

Braun Tasan, Murat, Matija Dreze, Miriam Barrios-Rodiles, Irma Lemmens, Haiyuan Yu, Julie M Sahalie, Ryan R Murray, Luba Roncari, Anne-Sophie de Smet, Kavitha Venkatesan, Jean-François Rual, Jean Vandenhaute, Michael E Cusick, Tony Pawson, David E Hill, Jan Tavernier, Jeffrey L Wrana, Frederick P Roth, and Marc Vidal (2009). "An experimentally derived confidence score for binary protein-protein interactions." In: *Nature Methods* 6.1, pp. 91–97. DOI: 10.1038/nmeth.1281.

Bulman, D E, S B Gangopadhyay, K G Bebchuck, R G Worton, and P N Ray (1991). "Point mutation in the human dystrophin gene: identification through western blot analysis." In: *Genomics* 10.2, pp. 457–460. DOI: 10.1016/0888-7543(91)90332-9.

Bystroff, Christopher and Anders Krogh (2008). "Hidden Markov Models for Prediction of Protein Features". In: *Methods in Molecular Biology*. Vol. 413. MIMB. Humana Press, pp. 173–198. DOI: 10.1007/978-1-59745-582-4_12.

Campbell, Iain Donald and Martin Baron (1991). "The structure and function of protein modules". In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 332.1263, pp. 199–203. ISSN: 0962-8436. DOI: 10.1098/rstb.1991.0045.

Chaisson Sanders, Ashley D, Xuefang Zhao, Ankit Malhotra, David Porubsky, Tobias Rausch, Eugene J Gardner, Oscar L Rodriguez, Li Guo, Ryan L Collins, Xian Fan, Jia Wen, Robert E Handsaker, Susan Fairley, Zev N Kronenberg, Xiangmeng Kong, Fereydoun Hormozdiari, Dillon Lee, Aaron M Wenger, Alex R Hastie, Danny Antaki, Thomas Anantharaman, Peter A Audano, Harrison Brand, Stuart Cantsilieris, Han Cao, Eliza Cerveira, Chong Chen, Xintong Chen, Chen-Shan Chin, Zechen Chong, Nelson T Chuang, Christine C Lambert, Deanna M

Church, Laura Clarke, Andrew Farrell, Joey Flores, Timur Galeev, David U Gorkin, Madhusudan Gujral, Victor Guryev, William Haynes Heaton, Jonas Korlach, Sushant Kumar, Jee Young Kwon, Ernest T Lam, Jong Eun Lee, Joyce Lee, Wan-Ping Lee, Sau Peng Lee, Shantao Li, Patrick Marks, Karine Viaud-Martinez, Sascha Meiers, Katherine M Munson, Fabio C P Navarro, Bradley J Nelson, Conor Nodzak, Amina Noor, Sofia Kyriazopoulou-Panagiotopoulou, Andy W C Pang, Yunjiang Qiu, Gabriel Rosanio, Mallory Ryan, Adrian Stütz, Diana C J Spierings, Alistair Ward, AnneMarie E Welch, Ming Xiao, Wei Xu, Chengsheng Zhang, Qihui Zhu, Xiangqun Zheng-Bradley, Ernesto Lowy, Sergei Yakneen, Steven McCarroll, Goo Jun, Li Ding, Chong Lek Koh, Bing Ren, Paul Flicek, Ken Chen, Mark B Gerstein, Pui-Yan Kwok, Peter M Lansdorp, Gabor T Marth, Jonathan Sebat, Xinghua Shi, Ali Bashir, Kai Ye, Scott E Devine, Michael E Talkowski, Ryan E Mills, Tobias Marschall, Jan O Korbel, Evan E Eichler, and Charles Lee (2019). "Multi-platform discovery of haplotype-resolved structural variation in human genomes." In: *Nature Communications* 10.1, p. 1784. ISSN: 2041-1723. DOI: 10.1038/s41467-018-08148-z.

Chen Li, S, Y Chen, P L Chen, Z D Sharp, and W H Lee (1996). "The nuclear localization sequences of the BRCA1 protein interact with the importin-alpha subunit of the nuclear transport signal receptor." In: *The Journal of Biological Chemistry* 271.51, pp. 32863–32868. DOI: 10.1074/jbc.271.51.32863.

Chen, Siwei, Robert Fragoza, Lambertus Klei, Yuan Liu, Jiebiao Wang, Kathryn Roeder, Bernie Devlin, and Haiyuan Yu (2018). "An interactome perturbation framework prioritizes damaging missense mutations for developmental disorders." In: *Nature Genetics* 50.7, pp. 1032–1040. ISSN: 1061-4036. DOI: 10.1038/s41588-018-0130-z.

Cheng, Jun, Guido Novati, Joshua Pan, Clare Bycroft, Akvilė Žemgulytė, Taylor Applebaum, Alexander Pritzel, Lai Hong Wong, Michal Zielinski, Tobias Sargeant, Rosalia G Schneider, Andrew W Senior, John Jumper, Demis Hassabis, Pushmeet Kohli, and Žiga Avsec (2023). "Accurate proteome-wide missense variant effect prediction with AlphaMissense." In: *Science* 381.6664, eadg7492. ISSN: 0036-8075. DOI: 10.1126/science.adg7492.

Chien, Bartel, Sternglanz, and Fields (1991). "The two-hybrid system: a method to identify and clone genes for proteins that interact with a protein of interest". In: *Proceedings of the National Academy of Sciences U S A* 88.21, pp. 9578–9582. DOI: 10.1073/pnas.88.21.9578.

Choi Olivet, Julien, Patricia Cassonnet, Pierre-Olivier Vidalain, Katja Luck, Luke Lambourne, Kerstin Spirohn, Irma Lemmens, Mélanie Dos Santos, Caroline Demeret, Louis Jones, Sudharshan Rangarajan, Wenting Bian, Eloi P Coutant, Yves L Janin, Sylvie van der Werf, Philipp Trepte, Erich E Wanker, Javier De Las Rivas, Jan Tavernier, Jean-Claude Twizere, Tong Hao, David E Hill, Marc

Vidal, Michael A Calderwood, and Yves Jacob (2019). "Maximizing binary interactome mapping with a minimal number of assays." In: *Nature Communications* 10.1, p. 3907. DOI: 10.1038/s41467-019-11809-2.

ClinVar Miner (2024). *ClinVar Miner*. Accessed: 2024-09-05.

Consortium, 1000 Genomes Project, Adam Auton, Lisa D Brooks, Richard M Durbin, Erik P Garrison, Hyun Min Kang, Jan O Korbel, Jonathan L Marchini, Shane McCarthy, Gil A McVean, and Gonçalo R Abecasis (2015). "A global reference for human genetic variation." In: *Nature* 526.7571, pp. 68–74. ISSN: 0028-0836. DOI: 10.1038/nature15393.

Copley, Richard R, Tobias Doerks, Ivica Letunic, and Peer Bork (2002). "Protein domain analysis in the era of complete genomes." In: *FEBS Letters* 513.1, pp. 129–134. DOI: 10.1016/s0014-5793(01)03289-6.

Davey, Norman E, M Madan Babu, Martin Blackledge, Alan Bridge, Salvador Capella-Gutierrez, Zsuzsanna Dosztanyi, Rachel Drysdale, Richard J Edwards, Arne Elofsson, Isabella C Felli, Toby J Gibson, Aleksandras Gutmanas, John M Hancock, Jen Harrow, Desmond Higgins, Cy M Jeffries, Philippe Le Mercier, Balint Mészáros, Marco Necci, Cedric Notredame, Sandra Orchard, Christos A Ouzounis, Rita Pancsa, Elena Papaleo, Roberta Pierattelli, Damiano Piovesan, Vasilis J Promponas, Patrick Ruch, Gabriella Rustici, Pedro Romero, Sirarat Sarntivijai, Gary Saunders, Benjamin Schuler, Malvika Sharan, Denis C Shields, Joel L Sussman, Jonathan A Tedds, Peter Tompa, Michael Turewicz, Jiri Vondrasek, Wim F Vranken, Bonnie Ann Wallace, Kanin Wichapong, and Silvio C E Tosatto (2019). "An intrinsically disordered proteins community for ELIXIR." In: *F1000Research* 8. DOI: 10.12688/f1000research.20136.1.

Davey, Norman E, Martha S Cyert, and Alan M Moses (2015). "Short linear motifs - ex nihilo evolution of protein regulation." In: *Cell Communication and Signaling* 13, p. 43. DOI: 10.1186/s12964-015-0120-z.

Davey, Norman E, Niall J Haslam, Denis C Shields, and Richard J Edwards (2011). "SLiMSearch 2.0: biological context for short linear motifs in proteins." In: *Nucleic Acids Research* 39.Web Server issue, W56–60. DOI: 10.1093/nar/gkr402.

Davey, Norman E, Kim Van Roey, Robert J Weatheritt, Grischa Toedt, Bora Uyar, Brigitte Altenberg, Aidan Budd, Francesca Diella, Holger Dinkel, and Toby J Gibson (2012). "Attributes of short linear motifs." In: *Molecular Biosystems* 8.1, pp. 268–281. DOI: 10.1039/c1mb05231d.

Dhanoa, Bajinder S, Tiziana Cogliati, Akhila G Satish, Elspeth A Bruford, and James S Friedman (2013). "Update on the Kelch-like (KLHL) gene family." In: *Human genomics* 7, p. 13. DOI: 10.1186/1479-7364-7-13.

Dill, Ken A and Justin L MacCallum (2012). "The protein-folding problem, 50 years on." In: *Science* 338.6110, pp. 1042–1046. DOI: 10.1126/science.1219021.

Ding Yuan, Fang, Priyadarshan K Damle, Larisa Litovchick, Ronny Drapkin, and Steven R Grossman (2020). "CtBP determines ovarian cancer cell fate through repression of death receptors." In: *Cell death & disease* 11.4, p. 286. DOI: 10.1038/s41419-020-2455-7.

Doolittle, Russell F. (1995). "The Multiplicity of Domains in Proteins". In: *Annual Review of Biochemistry* 64, pp. 287–314. DOI: 10.1146/annurev.bi.64.070195.001443.

Dosztányi, Peter Csizmok Tompa, and Simon (2005). "IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content." In: *Bioinformatics* 21.16, pp. 3433–3434. DOI: 10.1093/bioinformatics/bti541.

Dosztányi, Zsuzsanna (2018). "Prediction of protein disorder based on IUPred." In: *Protein Science* 27.1, pp. 331–340. DOI: 10.1002/pro.3334.

Dragulescu-Andrasi Chan, Carmel T, Abhijit De, Tarik F Massoud, and Sanjiv S Gambhir (2011). "Bioluminescence resonance energy transfer (BRET) imaging of protein-protein interactions within deep tissues of living subjects." In: *Proceedings of the National Academy of Sciences of the United States of America* 108.29, pp. 12060–12065. ISSN: 1091-6490. DOI: 10.1073/pnas.1100923108.

Dunker, A. Keith, Celeste J. Brown, and Zoran Obradovic (2002). "Identification and functions of usefully disordered proteins". In: *Advances in Protein Chemistry* 62, pp. 25–49. DOI: 10.1016/S0065-3233(02)62004-3.

Dyson Wright, Peter E (2005). "Intrinsically unstructured proteins and their functions." In: *Nature Reviews. Molecular Cell Biology* 6.3, pp. 197–208. DOI: 10.1038/nrm1589.

Edwards and Nicolas Palopoli (2014). "Computational Prediction of Short Linear Motifs from Protein Sequences". In: *Computational Peptidology*. Vol. 1268. Methods in Molecular Biology. Humana Press, pp. 89–141. DOI: 10.1007/978-1-4939-2285-7_5.

Felli, Isabella C. and Roberta Pierattelli (2015). *Intrinsically Disordered Proteins Studied by NMR Spectroscopy*. Springer. ISBN: 978-3-319-20197-9. DOI: 10.1007/978-3-319-20198-6.

Fields and Song (1989). "A novel genetic system to detect protein–protein interactions". In: *Nature* 340, pp. 245–246. DOI: 10.1038/340245a0.

Filograna De Tito, Stefano, Matteo Lo Monte, Rosario Oliva, Francesca Bruzzese, Maria Serena Roca, Antonella Zannetti, Adelaide Greco, Daniela Spano, Inmaculada Ayala, Assunta Liberti, Luigi Petraccone, Nina Dathan, Giuliana Catara, Laura Schembri, Antonino Colanzi, Alfredo Budillon, Andrea Rosario Beccari, Pompea Del Vecchio, Alberto Luini, Daniela Corda, and Carmen Valente (2024). "Identification and characterization of a new potent inhibitor tar-

geting CtBP1/BARS in melanoma cells." In: *Journal of Experimental & Clinical Cancer Research* 43.1, p. 137. DOI: 10.1186/s13046-024-03044-5.

Finn Mistry, Jaina, John Tate, Penny Coggill, Andreas Heger, Joanne E Pollington, O Luke Gavin, Prasad Gunasekaran, Goran Ceric, Kristoffer Forslund, Liisa Holm, Erik L L Sonnhammer, Sean R Eddy, and Alex Bateman (2010). "The Pfam protein families database." In: *Nucleic Acids Research* 38.Database issue, pp. D211–22. DOI: 10.1093/nar/gkp985.

Finn, Robert D, Alex Bateman, Jody Clements, Penelope Coggill, Ruth Y Eberhardt, Sean R Eddy, Andreas Heger, Kirstie Hetherington, Liisa Holm, Jaina Mistry, Erik L L Sonnhammer, John Tate, and Marco Punta (2014). "Pfam: the protein families database." In: *Nucleic Acids Research* 42.Database issue, pp. D222–30. DOI: 10.1093/nar/gkt1223.

Forbes Bindal, Nidhi, Sally Bamford, Charlotte Cole, Chai Yin Kok, David Beare, Mingming Jia, Rebecca Shepherd, Kenric Leung, Andrew Menzies, Jon W Teague, Peter J Campbell, Michael R Stratton, and P Andrew Futreal (2011). "COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer." In: *Nucleic Acids Research* 39.Database issue, pp. D945–50. DOI: 10.1093/nar/gkq929.

Fouassier, L, C C Yun, J G Fitz, and R B Doctor (2000). "Evidence for ezrin-radixin-moesin-binding phosphoprotein 50 (EBP50) self-association through PDZ-PDZ interactions." In: *The Journal of Biological Chemistry* 275.32, pp. 25039–25045. DOI: 10.1074/jbc.C000092200.

Fragoza, Robert, Jishnu Das, Shayne D Wierbowski, Jin Liang, Tina N Tran, Siqi Liang, Juan F Beltran, Christen A Rivera-Erick, Kaixiong Ye, Ting-Yi Wang, Li Yao, Matthew Mort, Peter D Stenson, David N Cooper, Xiaomu Wei, Alon Keinan, John C Schimenti, Andrew G Clark, and Haiyuan Yu (2019). "Extensive disruption of protein interactions by genetic variants across the allele frequency spectrum in human populations." In: *Nature Communications* 10.1, p. 4141. DOI: 10.1038/s41467-019-11959-3.

Freedman, M. H. and M. Sela (1966). "Recovery of antigenic activity upon reoxidation of completely reduced polyalanyl rabbit immunoglobulin G". In: *J. Biol. Chem.* 241.10, pp. 2383–2396.

Geist Lee, Chop Yan, Joelle Morgan Strom, José de Jesús Naveja, and Katja Luck (2024). "Generation of a high confidence set of domain-domain interface types to guide protein complex structure predictions by AlphaFold." In: *Bioinformatics*. DOI: 10.1093/bioinformatics/btae482.

Gilmore, T D (2006). "Introduction to NF-kappaB: players, pathways, perspectives." In: *Oncogene* 25.51, pp. 6680–6684. DOI: 10.1038/sj.onc.1209954.

Glover Williams, Lee (2004). "Interactions between BRCT repeats and phosphoproteins: tangled up in two." In: *Trends in Biochemical Sciences* 29.11, pp. 579–585. DOI: 10.1016/j.tibs.2004.09.010.

gnomAD (2024). *Genome Aggregation Database (gnomAD)*. Accessed: 2024-09-05.

Goh, Kwang-Il, Michael E Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási (2007). "The human disease network." In: *Proceedings of the National Academy of Sciences of the United States of America* 104.21, pp. 8685–8690. ISSN: 0027-8424. DOI: 10.1073/pnas.0701361104.

Gouw, Marc, Hugo Sámano-Sánchez, Kim Van Roey, Francesca Diella, Toby J Gibson, and Holger Dinkel (2017). "Exploring short linear motifs using the ELM database and tools." In: *Current Protocols in Bioinformatics* 58, pp. 8.22.1–8.22.35. DOI: 10.1002/cpbi.26.

Gouw, Hugo Sámano-Sánchez, Manjeet Kumar, András Zeke, Benjamin Lang, Benoit Bely, Lucía B Chemes, Norman E Davey, Ziqi Deng, Francesca Diella, Clara-Marie Gürth, Ann-Kathrin Huber, Stefan Kleinsorg, Lara S Schlegel, Nicolás Palopoli, Kim V Roey, Brigitte Altenberg, Attila Reményi, Holger Dinkel, and Toby J Gibson (2018). "The eukaryotic linear motif resource - 2018 update." In: *Nucleic Acids Research* 46.D1, pp. D428–D434. ISSN: 0305-1048. DOI: 10.1093/nar/gkx1077.

Goyet, Elise, Nathalie Bouquier, Vincent Ollendorff, and Julie Perroy (2016). "Fast and high resolution single-cell BRET imaging". In: *Scientific Reports* 6, Article 28231. DOI: 10.1038/srep28231.

Grozinger, C M and S L Schreiber (2000). "Regulation of histone deacetylase 4 and 5 and transcriptional activity by 14-3-3-dependent cellular localization." In: *Proceedings of the National Academy of Sciences of the United States of America* 97.14, pp. 7835–7840. DOI: 10.1073/pnas.140199597.

Grünberg, Raik, Julia V Burnier, Tony Ferrar, Violeta Beltran-Sastre, François Stricher, Almer M van der Sloot, Raquel Garcia-Olivas, Arrate Mallabiabarrena, Xavier Sanjuan, Timo Zimmermann, and Luis Serrano (2013). "Engineering of weak helper interactions for high-efficiency FRET probes". In: *Nature Methods* 10.10, pp. 1021–1027. DOI: 10.1038/nmeth.2625.

Gupta, Vandana A and Alan H Beggs (2014). "Kelch proteins: emerging roles in skeletal muscle development and diseases." In: *Skeletal muscle [electronic resource]* 4, p. 11. DOI: 10.1186/2044-5040-4-11.

Hall Unch, James, Brock F Binkowski, Michael P Valley, Braeden L Butler, Monika G Wood, Paul Otto, Kristopher Zimmerman, Gediminas Vidugiris, Thomas Machleidt, Matthew B Robers, Hélène A Benink, Christopher T Eggers, Michael R Slater, Poncho L Meisenheimer, Dieter H Klaubert, Frank Fan, Lance P Encell, and Keith V Wood (2012). "Engineered luciferase reporter from a deep sea

shrimp utilizing a novel imidazopyrazinone substrate." In: *ACS Chemical Biology* 7.11, pp. 1848–1857. DOI: 10.1021/cb3002478.

Hamosh Scott, Alan F, Joanna S Amberger, Carol A Bocchini, and Victor A McKusick (2005). "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders." In: *Nucleic Acids Research* 33.Database issue, pp. D514–7. DOI: 10.1093/nar/gki033.

Han, J.-D. J., N. Bertin, T. Hao, D. S. Goldberg, G. F. Berriz, L. V. Zhang, D. Dupuy, A. J. M. Walhout, M. E. Cusick, F. P. Roth, and M. Vidal (2004). *Evidence for dynamically organized modularity in the yeast protein–protein interaction network*. DOI: 10.1038/nature02654.

Harris, B Z and W A Lim (2001). "Mechanism and role of PDZ domains in signaling complex assembly." In: *Journal of Cell Science* 114.Pt 18, pp. 3219–3231. DOI: 10.1242/jcs.114.18.3219.

Hayden, Matthew S and Sankar Ghosh (2008). "Shared principles in NF-kappaB signaling." In: *Cell* 132.3, pp. 344–362. DOI: 10.1016/j.cell.2008.01.020.

Holmstrom, Erik D and David J Nesbitt (2016). "Biophysical Insights from Temperature-Dependent Single-Molecule Förster Resonance Energy Transfer." In: *Annual review of physical chemistry* 67, pp. 441–465. DOI: 10.1146/annurev-physchem-040215-112544.

Hsu, Lih-Ching (2007). "Identification and functional characterization of a PP1-binding site in BRCA1." In: *Biochemical and Biophysical Research Communications* 360.2, pp. 507–512. DOI: 10.1016/j.bbrc.2007.06.090.

Huttlin, Edward L., Raphael J. Bruckner, Joao A. Paulo, Joe R. Cannon, Lily Ting, Kurt Baltier, Greg Colby, Fana Gebreab, Melanie P. Gygi, Hannah Parzen, John Szpyt, Stanley Tam, Gabriela Zarraga, Laura Pontano-Vaites, Sharan Swarup, Anne E. White, Devin K. Schweppe, Ramin Rad, Brian K. Erickson, Robert A. Obar, K. G. Guruharsha, Kejie Li, Spyros Artavanis-Tsakonas, Steven P. Gygi, and J. Wade Harper (2017). "Architecture of the human interactome defines protein communities and disease networks". In: *Nature* 545.7655, pp. 505–509. ISSN: 0028-0836. DOI: 10.1038/nature22366.

Huttlin, Edward L., Richard J. Bruckner, Javier Navarrete-Perea, Jeffrey R. Cannon, Kevin Baltier, Fasil Gebreab, Martha P. Gygi, Austin Thornock, Genaro Zarraga, Shawn Tam, et al. (2021). "Dual proteome-scale networks reveal cell-specific remodeling of the human interactome". In: *Cell* 184.11, 3022–3040.e28. DOI: 10.1016/j.cell.2021.04.011.

Iakoucheva, Lilia M, Celeste J Brown, J David Lawson, Zoran Obradović, and A Keith Dunker (2002). "Intrinsic disorder in cell-signaling and cancer-associated proteins." In: *Journal of Molecular Biology* 323.3, pp. 573–584. ISSN: 0022-2836. DOI: 10.1016/s0022-2836(02)00969-5.

Idrees, Sobia and Keshav Raj Paudel (2024). "Proteome-wide assessment of human interactome as a source of capturing domain-motif and domain-domain interactions." In: *Journal of cell communication and signaling* 18.1, e12014. DOI: 10.1002/ccs3.12014.

Ingham Colwill, Karen, Caley Howard, Sabine Dettwiler, Caesar S H Lim, Joanna Yu, Kadija Hersi, Judith Raaijmakers, Gerald Gish, Geraldine Mbamalu, Lorne Taylor, Benny Yeung, Galina Vassilovski, Manish Amin, Fu Chen, Liudmila Matskova, Gösta Winberg, Ingemar Ernberg, Rune Linding, Paul O'donnell, Andrei Starostine, Walter Keller, Pavel Metalnikov, Chris Stark, and Tony Pawson (2005). "WW domains provide a platform for the assembly of multiprotein networks." In: *Molecular and Cellular Biology* 25.16, pp. 7092–7106. DOI: 10.1128/{MCB}.25.16.7092-7106.2005.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis (2021). "Highly accurate protein structure prediction with AlphaFold." In: *Nature* 596.7873, pp. 583–589. ISSN: 0028-0836. DOI: 10.1038/s41586-021-03819-2.

Karczewski Francioli, Laurent C, Grace Tiao, Beryl B Cummings, Jessica Alföldi, Qingbo Wang, Ryan L Collins, Kristen M Laricchia, Andrea Ganna, Daniel P Birnbaum, Laura D Gauthier, Harrison Brand, Matthew Solomonson, Nicholas A Watts, Daniel Rhodes, Moriel Singer-Berk, Eleina M England, Eleanor G Seaby, Jack A Kosmicki, Raymond K Walters, Katherine Tashman, Yossi Farjoun, Eric Banks, Timothy Poterba, Arcturus Wang, Cotton Seed, Nicola Whiffin, Jessica X Chong, Kaitlin E Samocha, Emma Pierce-Hoffman, Zachary Zappala, Anne H O'Donnell-Luria, Eric Vallabh Minikel, Ben Weisburd, Monkol Lek, James S Ware, Christopher Vittal, Irina M Armean, Louis Bergelson, Kristian Cibulskis, Kristen M Connolly, Miguel Covarrubias, Stacey Donnelly, Steven Ferriera, Stacey Gabriel, Jeff Gentry, Namrata Gupta, Thibault Jeandet, Diane Kaplan, Christopher Llanwarne, Ruchi Munshi, Sam Novod, Nikelle Petrillo, David Roazen, Valentin Ruano-Rubio, Andrea Saltzman, Molly Schleicher, Jose Soto, Kathleen Tibbetts, Charlotte Tolonen, Gordon Wade, Michael E Talkowski, Genome Aggregation Database Consortium, Benjamin M Neale, Mark J Daly, and Daniel G MacArthur (2020). "The mutational constraint spectrum quantified from variation in 141,456 humans." In: *Nature* 581.7809, pp. 434–443. ISSN: 0028-0836. DOI: 10.1038/s41586-020-2308-7.

Kim, Jiho and Regis Grailhe (2016). "Nanoluciferase signal brightness using furimazine substrates opens bioluminescence resonance energy transfer to widefield microscopy". In: *Cytometry Part A*. DOI: 10.1002/cyto.a.22870.

— (2024). "Nanoluciferase signal brightness using furimazine substrates opens bioluminescence resonance energy transfer to widefield microscopy". In: *Brief Report*. Free Access.

Kim, Mi-Sung, M Waseem Akhtar, Megumi Adachi, Melissa Mahgoub, Rhonda Bassel-Duby, Ege T Kavalali, Eric N Olson, and Lisa M Monteggia (2012). "An essential role for histone deacetylase 4 in synaptic plasticity and memory formation." In: *The Journal of Neuroscience* 32.32, pp. 10879–10886. DOI: 10.1523/{JNEUROSCI}.2089-12.2012.

Klug, Aaron (2010). "The discovery of zinc fingers and their applications in gene regulation and genome manipulation." In: *Annual Review of Biochemistry* 79, pp. 213–231. DOI: 10.1146/annurev-biochem-010909-095056.

Kobayashi, Hiroyuki, Louis-Philippe Picard, Anne-Marie Schönegge, and Michel Bouvier (2019). "Bioluminescence resonance energy transfer-based imaging of protein-protein interactions in living cells." In: *Nature Protocols* 14.4, pp. 1084–1107. DOI: 10.1038/s41596-019-0129-7.

Koipally Georgopoulos, K (2000). "Ikaros interactions with CtBP reveal a repression mechanism that is independent of histone deacetylase activity." In: *The Journal of Biological Chemistry* 275.26, pp. 19594–19602. DOI: 10.1074/jbc.M000254200.

Koonin, E V (1996). "Pseudouridine synthases: four families of enzymes containing a putative uridine-binding motif also conserved in dUTPases and dCTP deaminases." In: *Nucleic Acids Research* 24.12, pp. 2411–2415. DOI: 10.1093/nar/24.12.2411.

Kornau, H C, L T Schenker, M B Kennedy, and P H Seeburg (1995). "Domain interaction between NMDA receptor subunits and the postsynaptic density protein PSD-95." In: *Science* 269.5231, pp. 1737–1740. DOI: 10.1126/science.7569905.

Kumar, Manjeet, Sushama Michael, Jesús Alvarado-Valverde, András Zeke, Tamas Lazar, Juliana Glavina, Eszter Nagy-Kanta, Juan Mac Donagh, Zsofia E Kalman, Stefano Pascarelli, Nicolas Palopoli, László Dobson, Carmen Florencia Suarez, Kim Van Roey, Izabella Krystkowiak, Juan Esteban Griffin, Anurag Nagpal, Rajesh Bhardwaj, Francesca Diella, Bálint Mészáros, Kellie Dean, Norman E Davey, Rita Pancsa, Lucía B Chemes, and Toby J Gibson (2024). "ELM-the Eukaryotic Linear Motif resource-2024 update." In: *Nucleic Acids Research* 52.D1, pp. D442–D455. DOI: 10.1093/nar/gkad1058.

Lacoste, Jessica, Marzieh Haghighi, Shahan Haider, Zhen-Yuan Lin, Dmitri Segal, Chloe Reno, Wesley Wei Qian, Xueting Xiong, Hamdah Shafqat-Abbasi, Pearl V Ryder, Rebecca Senft, Beth A Cimini, Frederick P Roth, Michael Calderwood,

David Hill, Marc Vidal, S Stephen Yi, Nidhi Sahni, Jian Peng, Anne-Claude Gingras, Shantanu Singh, Anne E Carpenter, and Mikko Taipale (2023). "Pervasive mislocalization of pathogenic coding variants underlying human disorders." In: *BioRxiv*. DOI: 10.1101/2023.09.05.556368.

Landrum, Melissa J, Jennifer M Lee, Mark Benson, Garth Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Jeffrey Hoover, Wonhee Jang, Kenneth Katz, Michael Ovetsky, George Riley, Amanjeev Sethi, Ray Tully, Ricardo Villamarin-Salomon, Wendy Rubinstein, and Donna R Maglott (2016). "ClinVar: public archive of interpretations of clinically relevant variants." In: *Nucleic Acids Research* 44.D1, pp. D862–8. DOI: 10.1093/nar/gkv1222.

Lee (2010). "PDZ domains and their binding partners: structure, specificity, and modification." In: *Cell Communication and Signaling* 8, p. 8. DOI: 10.1186/1478-{811X}-8-8.

Lee, Hubrich, Varga, Christian Schäfer, Mareen Welzel, Eric Schumbera, Milena Djokic, Joelle M Strom, Jonas Schönfeld, Johanna L Geist, Feyza Polat, Toby J Gibson, Claudia Isabelle Keller Valsecchi, Manjeet Kumar, Ora Schueler-Furman, and Katja Luck (2024). "Systematic discovery of protein interaction interfaces using AlphaFold and experimental validation." In: *Molecular Systems Biology* 20.2, pp. 75–97. ISSN: 1744-4292. DOI: 10.1038/s44320-023-00005-6.

Lee, Olzmann, Lih-Shen Chin, and Lian Li (2011). "Mutations associated with Charcot-Marie-Tooth disease cause SIMPLE protein mislocalization and degradation by the proteasome and aggresome-autophagy pathways." In: *Journal of Cell Science* 124.Pt 19, pp. 3319–3331. DOI: 10.1242/jcs.087114.

Lek, Monkol, Konrad J Karczewski, Eric V Minikel, Kaitlin E Samocha, Eric Banks, Timothy Fennell, Anne H O'Donnell-Luria, James S Ware, Andrew J Hill, Beryl B Cummings, Taru Tukiainen, Daniel P Birnbaum, Jack A Kosmicki, Laramie E Duncan, Karol Estrada, Fengmei Zhao, James Zou, Emma Pierce-Hoffman, Joanne Berghout, David N Cooper, Nicole Deflaux, Mark DePristo, Ron Do, Jason Flannick, Menachem Fromer, Laura Gauthier, Jackie Goldstein, Namrata Gupta, Daniel Howrigan, Adam Kiezun, Mitja I Kurki, Ami Levy Moonshine, Pradeep Natarajan, Lorena Orozco, Gina M Peloso, Ryan Poplin, Manuel A Rivas, Valentin Ruano-Rubio, Samuel A Rose, Douglas M Ruderfer, Khalid Shakir, Peter D Stenson, Christine Stevens, Brett P Thomas, Grace Tiao, Maria T Tusie-Luna, Ben Weisburd, Hong-Hee Won, Dongmei Yu, David M Altshuler, Diego Ardissino, Michael Boehnke, John Danesh, Stacey Donnelly, Roberto Elosua, Jose C Florez, Stacey B Gabriel, Gad Getz, Stephen J Glatt, Christina M Hultman, Sekar Kathiresan, Markku Laakso, Steven McCarroll, Mark I McCarthy, Dermot McGovern, Ruth McPherson, Benjamin M Neale, Aarno Palotie, Shaun M Purcell, Danish Saleheen, Jeremiah M Scharf, Pamela Sklar, Patrick F Sul-

livan, Jaakko Tuomilehto, Ming T Tsuang, Hugh C Watkins, James G Wilson, Mark J Daly, Daniel G MacArthur, and Exome Aggregation Consortium (2016). "Analysis of protein-coding genetic variation in 60,706 humans." In: *Nature* 536.7616, pp. 285–291. ISSN: 0028-0836. DOI: 10.1038/nature19057.

Letunic, Ivica, Supriya Khedkar, and Peer Bork (2021). "SMART: recent updates, new developments and status in 2020." In: *Nucleic Acids Research* 49.D1, pp. D458–D460. DOI: 10.1093/nar/gkaa937.

Li Wang, Fei, Qiao Wang, Na Zhang, Jumei Zheng, Maiqing Zheng, Ranran Liu, Huanxian Cui, Jie Wen, and Guiping Zhao (2020). "SPOP promotes ubiquitination and degradation of MyD88 to suppress the innate immune response." In: *PLoS Pathogens* 16.5, e1008188. DOI: 10.1371/journal.ppat.1008188.

Lievens, Peelman, De Bosscher, Lemmens, and Jan Tavernier (2011). "MAPPIT: a protein interaction toolbox built on insights in cytokine receptor signaling". In: *Cytokine Growth Factor Reviews* 22.5-6, pp. 321–329. DOI: 10.1016/j.cytogfr.2011.11.001.

Lin Smith, Edwin R, Hidehisa Takahashi, Ka Chun Lai, Skylar Martin-Brown, Laurence Florens, Michael P Washburn, Joan W Conaway, Ronald C Conaway, and Ali Shilatifard (2010). "AFF4, a component of the ELL/P-TEFb elongation complex and a shared subunit of MLL chimeras, can link transcription elongation to leukemia." In: *Molecular Cell* 37.3, pp. 429–437. ISSN: 1097-4164. DOI: 10.1016/j.molcel.2010.01.026.

Livesey, Benjamin J and Joseph A Marsh (2022). "Interpreting protein variant effects with computational predictors and deep mutational scanning." In: *Disease Models & Mechanisms* 15.6. DOI: 10.1242/dmm.049510.

Luck, Katja, Sebastian Charbonnier, and Gilles Travé (2012). "The emerging contribution of sequence context to the specificity of protein interactions mediated by PDZ domains". In: *FEBS Letters* 586.17, pp. 2648–2661. DOI: 10.1016/j.febslet.2012.03.056.

Luck, Katja, Dae-Kyum Kim, Luke Lambourne, Kerstin Spirohn, Bridget E Begg, Wenting Bian, Ruth Brignall, Tiziana Cafarelli, Francisco J Campos-Laborie, Benoit Charloteaux, Dongsic Choi, Atina G Coté, Meaghan Daley, Steven Deimling, Alice Desbuleux, Amélie Dricot, Marinella Gebbia, Madeleine F Hardy, Nishka Kishore, Jennifer J Knapp, István A Kovács, Irma Lemmens, Miles W Mee, Joseph C Mellor, Carl Pollis, Carles Pons, Aaron D Richardson, Sadie Schlabach, Bridget Teeking, Anupama Yadav, Mariana Babor, Dawit Balcha, Omer Basha, Christian Bowman-Colin, Suet-Feung Chin, Soon Gang Choi, Claudia Colabella, Georges Coppin, Cassandra D'Amata, David De Ridder, Steffi De Rouck, Miquel Duran-Frigola, Hanane Ennajdaoui, Florian Goebels, Liana Goehring, Anjali Gopal, Ghazal Haddad, Elodie Hatchi, Mohamed Helmy, Yves Jacob, Yoseph Kassa, Serena Landini, Roujia Li, Natascha van Lieshout, An-

drew MacWilliams, Dylan Markey, Joseph N Paulson, Sudharshan Rangarajan, John Rasla, Ashyad Rayhan, Thomas Rolland, Adriana San-Miguel, Yun Shen, Dayag Sheykhkarimli, Gloria M Sheynkman, Eyal Simonovsky, Murat Taşan, Alexander Tejeda, Vincent Tropepe, Jean-Claude Twizere, Yang Wang, Robert J Weatheritt, Jochen Weile, Yu Xia, Xinping Yang, Esti Yeger-Lotem, Quan Zhong, Patrick Aloy, Gary D Bader, Javier De Las Rivas, Suzanne Gaudet, Tong Hao, Janusz Rak, Jan Tavernier, David E Hill, Marc Vidal, Frederick P Roth, and Michael A Calderwood (2020). "A reference map of the human binary protein interactome." In: *Nature* 580.7803, pp. 402–408. ISSN: 0028-0836. DOI: 10.1038/s41586-020-2188-x.

Ludes-Meyers Kil, Hyunsuk, Andrzej K Bednarek, Jeff Drake, Mark T Bedford, and C Marcelo Aldaz (2004). "WWOX binds the specific proline-rich ligand PPXY: identification of candidate interacting proteins." In: *Oncogene* 23.29, pp. 5049–5055. DOI: 10.1038/sj.onc.1207680.

Luo Lin, Chengqi, Erin Guest, Alexander S Garrett, Nima Mohaghegh, Selene Swanson, Stacy Marshall, Laurence Florens, Michael P Washburn, and Ali Shilatifard (2012). "The super elongation complex family of RNA polymerase II elongation factors: gene target specificity and transcriptional output." In: *Molecular and Cellular Biology* 32.13, pp. 2608–2617. DOI: 10.1128/{MCB}.00182-12.

Luo, X., Q. He, Y. Huang, and M. S. Sheikh (2005). "Cloning and characterization of a p53 and DNA damage down-regulated gene PIQ that codes for a novel calmodulin-binding IQ motif protein and is up-regulated in gastrointestinal cancers". In: *Cancer Research* 65, pp. 10725–10733.

Martino, Elisa, Sara Chiarugi, Francesco Margheriti, and Gianpiero Garau (2021). "Mapping, structure and modulation of PPI". In: *Frontiers in Chemistry* 9, p. 718405. DOI: 10.3389/fchem.2021.718405.

Melhuish Wotton, D (2000). "The interaction of the carboxyl terminus-binding protein with the Smad corepressor TGIF is disrupted by a holoprosencephaly mutation in TGIF." In: *The Journal of Biological Chemistry* 275.50, pp. 39762–39766. DOI: 10.1074/jbc.C000416200.

Mészáros, Bálint, István Simon, and Zsuzsanna Dosztányi (2009). "Prediction of protein binding regions in disordered proteins." In: *PLoS Computational Biology* 5.5, e1000376. DOI: 10.1371/journal.pcbi.1000376.

Meyer Kirchner, Marieluise, Bora Uyar, Jing-Yuan Cheng, Giulia Russo, Luis R Hernandez-Miranda, Anna Szymborska, Henrik Zauber, Ina-Maria Rudolph, Thomas E Willnow, Altuna Akalin, Volker Haucke, Holger Gerhardt, Carmen Birchmeier, Ralf Kühn, Michael Krauss, Sebastian Diecke, Juan M Pascual, and Matthias Selbach (2018). "Mutations in disordered regions can cause disease by creating dileucine motifs." In: *Cell* 175.1, 239–253.e17. ISSN: 00928674. DOI: 10.1016/j.cell.2018.08.019.

Mihalič, Filip, Leandro Simonetti, Girolamo Giudice, Marie Rubin Sander, Richard Lindqvist, Marie Berit Akpiroro Peters, Caroline Benz, Eszter Kassa, Dilip Badgujar, Raviteja Inturi, Muhammad Ali, Izabella Krystkowiak, Ahmed Sayadi, Eva Andersson, Hanna Aronsson, Ola Söderberg, Doreen Dobritzsch, Evangelia Petsalaki, Anna K Överby, Per Jemth, Norman E Davey, and Ylva Ivarsson (2023). "Large-scale phage-based screening reveals extensive pan-viral mimicry of host short linear motifs." In: *Nature Communications* 14.1, p. 2409. DOI: 10.1038/s41467-023-38015-5.

Mosca, Roberto, Arnaud Céol, Amelie Stein, Roger Olivella, and Patrick Aloy (2014). "3did: a catalog of domain-based interactions of known three-dimensional structure". In: *Nucleic Acids Research* 42.Database issue, pp. D374–D379. DOI: 10.1093/nar/gkt887. eprint: 2013Sep29.

Nesta, Alex V, Denisse Tafur, and Christine R Beck (2021). "Hotspots of human mutation." In: *Trends in Genetics* 37.8, pp. 717–729. ISSN: 01689525. DOI: 10.1016/j.tig.2020.10.003.

Nooren Thornton, Janet M. (2003). "Diversity of protein–protein interactions". In: *The EMBO Journal* 22.14, pp. 3486–3492. DOI: 10.1093/emboj/cdg359.

Northrop, J. H. (1930). "CRYSTALLINE PEPSIN: I. ISOLATION AND TESTS OF PURITY". In: *The Journal of General Physiology* 13.6, pp. 739–766. DOI: 10.1085/jgp.13.6.739.

Oldfield, Christopher J and A Keith Dunker (2014). "Intrinsically disordered proteins and intrinsically disordered protein regions." In: *Annual Review of Biochemistry* 83, pp. 553–584. DOI: 10.1146/annurev-biochem-072711-164947.

Oliver Bitoun, Emmanuelle, Joanne Clark, Emma L Jones, and Kay E Davies (2004). "Mediation of Af4 protein function in the cerebellum by Siah proteins." In: *Proceedings of the National Academy of Sciences of the United States of America* 101.41, pp. 14901–14906. DOI: 10.1073/pnas.0406196101.

Oxley Anthis, Nicholas J, Edward D Lowe, Ioannis Vakonakis, Iain D Campbell, and Kate L Wegener (2008). "An integrin phosphorylation switch: the effect of beta3 integrin tail phosphorylation on Dok1 and talin binding." In: *The Journal of Biological Chemistry* 283.9, pp. 5420–5426. DOI: 10.1074/jbc.M709435200.

Paysan-Lafosse, Typhaine, Matthias Blum, Sara Chuguransky, Tiago Grego, Beatriz Lázaro Pinto, Gustavo A Salazar, Maxwell L Bileschi, Peer Bork, Alan Bridge, Lucy Colwell, Julian Gough, Daniel H Haft, Ivica Letunić, Aron Marchler-Bauer, Huaiyu Mi, Darren A Natale, Christine A Orengo, Arun P Pandurangan, Catherine Rivoire, Christian J A Sigrist, Ian Sillitoe, Narmada Thanki, Paul D Thomas, Silvio C E Tosatto, Cathy H Wu, and Alex Bateman (2023). "InterPro in 2022." In: *Nucleic Acids Research* 51.D1, pp. D418–D427. DOI: 10.1093/nar/gkac993.

Peng, Zhenling, Marcin J Mizianty, Bin Xue, Lukasz Kurgan, and Vladimir N Uversky (2012). "More than just tails: intrinsic disorder in histone proteins." In: *Molecular Biosystems* 8.7, pp. 1886–1901. DOI: 10.1039/c2mb25102g.

Pennington, K L, T Y Chan, M P Torres, and J L Andersen (2018). "The dynamic and stress-adaptive signaling hub of 14-3-3: emerging mechanisms of regulation and context-dependent protein-protein interactions." In: *Oncogene* 37.42, pp. 5587–5604. DOI: 10.1038/s41388-018-0348-3.

Petsalaki Stark, Alexander, Eduardo García-Urdiales, and Robert B Russell (2009). "Accurate prediction of peptide binding sites on protein surfaces." In: *PLoS Computational Biology* 5.3, e1000335. DOI: 10.1371/journal.pcbi.1000335.

Pfleger Seeber, Ruth M and Karin A Eidne (2006). "Bioluminescence resonance energy transfer (BRET) for the real-time detection of protein-protein interactions." In: *Nature Protocols* 1.1, pp. 337–345. DOI: 10.1038/nprot.2006.52.

Pierce, Michael M., C. S. Raman, and Barry T. Nall (1999). "Isothermal Titration Calorimetry of Protein–Protein Interactions". In: *Methods* 19.2, pp. 213–221. DOI: 10.1016/S1046-2023(99)00009-0.

Puntervoll Linding, Rune, Christine Gemünd, Sophie Chabanis-Davidson, Morten Mattingsdal, Scott Cameron, David M A Martin, Gabriele Ausiello, Barbara Brannetti, Anna Costantini, Fabrizio Ferrè, Vincenza Maselli, Allegra Via, Gianni Cesareni, Francesca Diella, Giulio Superti-Furga, Lucjan Wyrwicz, Chenna Ramu, Caroline McGuigan, Rambabu Gudavalli, Ivica Letunic, Peer Bork, Leszek Rychlewski, Bernhard Küster, Manuela Helmer-Citterich, William N Hunter, Rein Aasland, and Toby J Gibson (2003). "ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins." In: *Nucleic Acids Research* 31.13, pp. 3625–3630. DOI: 10.1093/nar/gkg545.

Ramirez-Martinez, Andres, Bercin Kutluk Cenik, Svetlana Bezprozvannaya, Beibei Chen, Rhonda Bassel-Duby, Ning Liu, and Eric N Olson (2017). "KLHL41 stabilizes skeletal muscle sarcomeres by nonproteolytic ubiquitination." In: *eLife* 6. DOI: 10.7554/{eLife}.26439.

Rodríguez, J A and B R Henderson (2000). "Identification of a functional nuclear export sequence in BRCA1." In: *The Journal of Biological Chemistry* 275.49, pp. 38589–38596. DOI: 10.1074/jbc.M003851200.

Rolland, Thomas, Murat Taşan, Benoit Charloteaux, Samuel J Pevzner, Quan Zhong, Nidhi Sahni, Song Yi, Irma Lemmens, Celia Fontanillo, Roberto Mosca, Atanas Kamburov, Susan D Ghiassian, Xinping Yang, Lila Ghamsari, Dawit Balcha, Bridget E Begg, Pascal Braun, Marc Brehme, Martin P Broly, Anne-Ruxandra Carvunis, Dan Convery-Zupan, Roser Corominas, Jasmin Coulombe-Huntington, Elizabeth Dann, Matija Dreze, Amélie Dricot, Changyu Fan, Eric Franzosa, Fana Gebreab, Bryan J Gutierrez, Madeleine F Hardy, Mike Jin, Shuli Kang, Ruth Kiros, Guan Ning Lin, Katja Luck, Andrew MacWilliams, Jörg

Menche, Ryan R Murray, Alexandre Palagi, Matthew M Poulin, Xavier Rambout, John Rasla, Patrick Reichert, Viviana Romero, Elien Ruyssinck, Julie M Sahalie, Annemarie Scholz, Akash A Shah, Amitabh Sharma, Yun Shen, Kerstin Spirohn, Stanley Tam, Alexander O Tejeda, Shelly A Wanamaker, Jean-Claude Twizere, Kerwin Vega, Jennifer Walsh, Michael E Cusick, Yu Xia, Albert-László Barabási, Lilia M Iakoucheva, Patrick Aloy, Javier De Las Rivas, Jan Tavernier, Michael A Calderwood, David E Hill, Tong Hao, Frederick P Roth, and Marc Vidal (2014). "A proteome-scale map of the human interactome network." In: *Cell* 159.5, pp. 1212–1226. DOI: 10.1016/j.cell.2014.10.050.

Sahni, Nidhi, Song Yi, Mikko Taipale, Juan I Fuxman Bass, Jasmin Coulombe-Huntington, Fan Yang, Jian Peng, Jochen Weile, Georgios I Karras, Yang Wang, István A Kovács, Atanas Kamburov, Irina Krykbaeva, Mandy H Lam, George Tucker, Vikram Khurana, Amitabh Sharma, Yang-Yu Liu, Nozomu Yachie, Quan Zhong, Yun Shen, Alexandre Palagi, Adriana San-Miguel, Changyu Fan, Dawit Balcha, Amelie Dricot, Daniel M Jordan, Jennifer M Walsh, Akash A Shah, Xinping Yang, Ani K Stoyanova, Alex Leighton, Michael A Calderwood, Yves Jacob, Michael E Cusick, Kourosh Salehi-Ashtiani, Luke J Whitesell, Shamil Sunyaev, Bonnie Berger, Albert-László Barabási, Benoit Charloteaux, David E Hill, Tong Hao, Frederick P Roth, Yu Xia, Albertha J M Walhout, Susan Lindquist, and Marc Vidal (2015). "Widespread macromolecular interaction perturbations in human genetic disorders." In: *Cell* 161.3, pp. 647–660. DOI: 10.1016/j.cell.2015.04.013.

Sahni, Nidhi, Song Yi, Quan Zhong, Noor Jailkhani, Benoit Charloteaux, Michael E Cusick, and Marc Vidal (2013). "Edgotype: a fundamental link between genotype and phenotype." In: *Current Opinion in Genetics & Development* 23.6, pp. 649–657. DOI: 10.1016/j.gde.2013.11.002.

Santelli Leone, Marilisa, Chenlong Li, Toru Fukushima, Nicholas E Preece, Arthur J Olson, Kathryn R Ely, John C Reed, Maurizio Pellecchia, Robert C Liddington, and Shu-ichi Matsuzawa (2005). "Structural analysis of Siah1-Siah-interacting protein interactions and insights into the assembly of an E3 ligase multiprotein complex." In: *The Journal of Biological Chemistry* 280.40, pp. 34278–34287. DOI: 10.1074/jbc.M506707200.

Schreiber, G, G Haran, and H-X Zhou (2009). "Fundamental aspects of protein-protein association kinetics." In: *Chemical Reviews* 109.3, pp. 839–860. DOI: 10.1021/cr800373w.

Schultz, J., F. Milpetz, P. Bork, and C.P. Ponting (1998). "SMART, a simple modular architecture research tool: identification of signaling domains". In: *Proceedings of the National Academy of Sciences U.S.A.* 95.11, pp. 5857–5864. DOI: 10.1073/pnas.95.11.5857.

Sekar, Rajesh Babu and Ammasi Periasamy (2003). "Fluorescence resonance energy transfer (FRET) microscopy imaging of live cell protein localizations". In: *Journal of Cell Biology* 160.5, pp. 629–633. DOI: 10.1083/jcb.200210140.

Shaner Lambert, Gerard G., Andrew Chammas, Yuhui Ni, Paula J. Cranfill, Michelle A. Baird, Brittney R. Sell, John R. Allen, Richard N. Day, Maria Israelsson, Michael W. Davidson, and Jiwu Wang (2013). "A bright monomeric green fluorescent protein derived from Branchiostoma lanceolatum". In: *Nature Methods* 10, pp. 407–409. DOI: 10.1038/nmeth.2413.

Starita, Lea M, Muhtadi M Islam, Tapahsama Banerjee, Aleksandra I Adamovich, Justin Gullingsrud, Stanley Fields, Jay Shendure, and Jeffrey D Parvin (2018). "A Multiplex Homology-Directed DNA Repair Assay Reveals the Impact of More Than 1,000 BRCA1 Missense Substitution Variants on Protein Function." In: *American Journal of Human Genetics* 103.4, pp. 498–508. ISSN: 00029297. DOI: 10.1016/j.ajhg.2018.07.016.

Stogios, Peter J and Gilbert G Privé (2004). "The BACK domain in BTB-kelch proteins." In: *Trends in Biochemical Sciences* 29.12, pp. 634–637. DOI: 10.1016/j.tibs.2004.10.003.

Sunyaev, S R, F Eisenhaber, I V Rodchenkov, B Eisenhaber, V G Tumanyan, and E N Kuznetsov (1999). "PSIC: profile extraction from sequence alignments with position-specific counts of independent observations." In: *Protein Engineering* 12.5, pp. 387–394. DOI: 10.1093/protein/12.5.387.

Tadokoro Shattil, Sanford J, Koji Eto, Vera Tai, Robert C Liddington, Jose M de Pereda, Mark H Ginsberg, and David A Calderwood (2003). "Talin binding to integrin beta tails: a final common step in integrin activation." In: *Science* 302.5642, pp. 103–106. ISSN: 1095-9203. DOI: 10.1126/science.1086652.

Taniguchi, Koji and Michael Karin (2018). "NF-, inflammation, immunity and cancer: coming of age." In: *Nature Reviews. Immunology* 18.5, pp. 309–324. DOI: 10.1038/nri.2017.142.

Tompa, Peter (2002). "Intrinsically unstructured proteins." In: *Trends in Biochemical Sciences* 27.10, pp. 527–533. DOI: 10.1016/s0968-0004(02)02169-2.

— (2011). "Unstructural biology coming of age." In: *Current Opinion in Structural Biology* 21.3, pp. 419–425. DOI: 10.1016/j.sbi.2011.03.012.

— (2012). "Intrinsically disordered proteins: a 10-year recap". In: *Trends in Biochemical Sciences* 37.12. Available at: ptompa@vub.ac.be, pp. 509–516. DOI: 10.1016/j.tibs.2012.08.009.

Tompa, Peter, Norman E Davey, Toby J Gibson, and M Madan Babu (2014). "A million peptide motifs for the molecular biologist." In: *Molecular Cell* 55.2, pp. 161–169. DOI: 10.1016/j.molcel.2014.05.032.

Trepte Secker, Christopher, Soon Gang Choi, Julien Olivet, Eduardo Silva Ramos, Patricia Cassonnet, Sabrina Golusik, Martina Zenkner, Stephanie Beetz, Marcel

Sperling, Yang Wang, Tong Hao, Kerstin Spirohn, Jean-Claude Twizere, Michael A. Calderwood, David E. Hill, Yves Jacob, Marc Vidal, and Erich E. Wanker (2021). "A quantitative mapping approach to identify direct interactions within complexomes". In: *BioRxiv*. DOI: 10.1101/2021.08.25.457734.

Trepte, Kruse, Kostova, Hoffmann, Buntru, Tempelmeier, Secker, Diez, Schulz, Klockmeier, Zenkner, Golusik, Rau, Schnoegl, Garner, and Erich Wanker (2018). "LuTHy: a double-readout bioluminescence-based two-hybrid technology for quantitative mapping of protein-protein interactions in mammalian cells." In: *Molecular Systems Biology* 14.7, e8071. DOI: 10.15252/msb.20178071.

Uversky (2014). *Intrinsically Disordered Proteins*. Switzerland: Springer International Publishing, pp. XV, 61. ISBN: 978-3-319-08920-1.

Uversky, Christopher J Oldfield, and A Keith Dunker (2005). "Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling." In: *Journal of Molecular Recognition* 18.5, pp. 343–384. DOI: 10.1002/jmr.747.

Uyar, Bora, Robert J Weatheritt, Holger Dinkel, Norman E Davey, and Toby J Gibson (2014). "Proteome-wide analysis of human disease mutations in short linear motifs: neglected players in cancer?" In: *Molecular Biosystems* 10.10, pp. 2626–2642. DOI: 10.1039/c4mb00290c.

Valente Luini, Alberto and Daniela Corda (2013). "Components of the CtBP1/BARS-dependent fission machinery." In: *Histochemistry and Cell Biology* 140.4, pp. 407–421. DOI: 10.1007/s00418-013-1138-1.

Van Roey, Kim, Toby J Gibson, and Norman E Davey (2012). "Motif switches: decision-making in cell regulation." In: *Current Opinion in Structural Biology* 22.3, pp. 378–385. DOI: 10.1016/j.sbi.2012.03.004.

Van Roey, Kim, Bora Uyar, Robert J Weatheritt, Holger Dinkel, Markus Seiler, Aidan Budd, Toby J Gibson, and Norman E Davey (2014). "Short linear motifs: ubiquitous and functionally diverse protein interaction modules directing cell regulation." In: *Chemical Reviews* 114.13, pp. 6733–6778. DOI: 10.1021/cr400585q.

Velthuis, Aartjan J W te, Philippe A Sakalis, Donald A Fowler, and Christoph P Bagowski (2011). "Genome-wide analysis of PDZ domain binding reveals inherent functional overlap within the PDZ interaction network." In: *Plos One* 6.1, e16047. DOI: 10.1371/journal.pone.0016047.

Vidal, Marc, Michael E Cusick, and Albert-László Barabási (2011). "Interactome networks and human disease." In: *Cell* 144.6, pp. 986–998. ISSN: 1097-4172. DOI: 10.1016/j.cell.2011.02.016.

Visscher, Peter M, Matthew A Brown, Mark I McCarthy, and Jian Yang (2012). "Five years of GWAS discovery." In: *American Journal of Human Genetics* 90.1, pp. 7–24. DOI: 10.1016/j.ajhg.2011.11.029.

Vogel, Christine, Carlo Berzuini, Matthew Bashton, Julian Gough, and Sarah A. Teichmann (Year). "Supra-domains: Evolutionary units larger than single protein domains". In: *Journal Name* Volume.Issue, pages. DOI: 10.XXXX/XXXXXX.

Vogel, Steven S, Christopher Thaler, and Srinagesh V Koushik (2006). "Fanciful FRET". In: *Science's STKE* 2006.331, re2. DOI: 10.1126/stke.3312006re2.

Wakeling, Emma, Meriel McEntagart, Michael Bruccoleri, Charles Shaw-Smith, Karen L Stals, Matthew Wakeling, Angela Barnicoat, Clare Beesley, DDD Study, Andrea K Hanson-Kahn, Mary Kukolich, David A Stevenson, Philippe M Campeau, Sian Ellard, Sarah H Elsea, Xiang-Jiao Yang, and Richard C Caswell (2021). "Missense substitutions at a conserved 14-3-3 binding site in HDAC4 cause a novel intellectual disability syndrome." In: *HGG advances* 2.1, p. 100015. DOI: 10.1016/j.xhgg.2020.100015.

Wang, Jia Chen, and Mingjie Zhang (2010). "Extensions of PDZ domains as important structural and functional elements." In: *Protein & cell* 1.8, pp. 737–751. DOI: 10.1007/s13238-010-0099-6.

Wang, Jiyao, Farideh Chitsaz, Myra K Derbyshire, Noreen R Gonzales, Marc Gwadz, Shennan Lu, Gabriele H Marchler, James S Song, Narmada Thanki, Roxanne A Yamashita, Mingzhang Yang, Dachuan Zhang, Chanjuan Zheng, Christopher J Lanczycki, and Aron Marchler-Bauer (2023). "The conserved domain database in 2023." In: *Nucleic Acids Research* 51.D1, pp. D384–D388. DOI: 10.1093/nar/gkac1096.

Wang Kruhlak, M J, J Wu, N R Bertos, M Vezmar, B I Posner, D P Bazett-Jones, and X J Yang (2000). "Regulation of histone deacetylase 4 by binding of 14-3-3 proteins." In: *Molecular and Cellular Biology* 20.18, pp. 6904–6912. DOI: 10.1128/{MCB}.20.18.6904-6912.2000.

Weatheritt, Robert J and Toby J Gibson (2012). "Linear motifs: lost in (pre)translation." In: *Trends in Biochemical Sciences* 37.8, pp. 333–341. DOI: 10.1016/j.tibs.2012.05.001.

Wegener Partridge, Anthony W, Jaewon Han, Andrew R Pickford, Robert C Liddington, Mark H Ginsberg, and Iain D Campbell (2007). "Structural basis of integrin activation by talin." In: *Cell* 128.1, pp. 171–182. ISSN: 0092-8674. DOI: 10.1016/j.cell.2006.10.048.

Wierbowski, Shayne D., Robert Fragoza, Siqi Liang, and Haiyuan Yu (2018). "Extracting complementary insights from molecular phenotypes for prioritization of disease-associated mutations". In: *Current Opinion in Systems Biology* 11, pp. 107–116. ISSN: 24523100. DOI: 10.1016/j.coisb.2018.09.006.

Williams, R S, R Green, and J N Glover (2001). "Crystal structure of the BRCT repeat region from the breast cancer-associated protein BRCA1." In: *Nature Structural Biology* 8.10, pp. 838–842. ISSN: 1072-8368. DOI: 10.1038/nsb1001-838.

Wilson, Carter J, Wing-Yiu Choy, and Mikko Karttunen (2022). "Alphafold2: A role for disordered protein/region prediction?" In: *International Journal of Molecular Sciences* 23.9, p. 4591. DOI: 10.3390/ijms23094591.

Wright, Dyson (1999). "Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm." In: *Journal of Molecular Biology* 293.2, pp. 321–331. DOI: 10.1006/jmbi.1999.3110.

— (2015). "Intrinsically disordered proteins in cellular signalling and regulation." In: *Nature Reviews. Molecular Cell Biology* 16.1, pp. 18–29. DOI: 10.1038/nrm3920.

Xu Piston, D W and C H Johnson (1999). "A bioluminescence resonance energy transfer (BRET) system: application to interacting circadian clock proteins." In: *Proceedings of the National Academy of Sciences of the United States of America* 96.1, pp. 151–156. DOI: 10.1073/pnas.96.1.151.

Yuen, Michaela and Coen A C Ottenheijm (2020). "Nebulin: big protein with big responsibilities." In: *Journal of muscle research and cell motility* 41.1, pp. 103–124. DOI: 10.1007/s10974-019-09565-3.

Zhang, Yingnan, Brent A Appleton, Christian Wiesmann, Ted Lau, Mike Costa, Rami N Hannoush, and Sachdev S Sidhu (2009). "Inhibition of Wnt signaling by Dishevelled PDZ peptides." In: *Nature Chemical Biology* 5.4, pp. 217–219. DOI: 10.1038/nchembio.152.

Zhong, Quan, Nicolas Simonis, Qian-Ru Li, Benoit Charloteaux, Fabien Heuze, Niels Klitgord, Stanley Tam, Haiyuan Yu, Kavitha Venkatesan, Danny Mou, Venus Swearingen, Muhammed A Yildirim, Han Yan, Amélie Dricot, David Szeto, Chenwei Lin, Tong Hao, Changyu Fan, Stuart Milstein, Denis Dupuy, Robert Brasseur, David E Hill, Michael E Cusick, and Marc Vidal (2009). "Edgetic perturbation models of human inherited disorders." In: *Molecular Systems Biology* 5, p. 321. DOI: 10.1038/msb.2009.80.

Zhou, Huan-Xiang (2012). "Intrinsic disorder: signaling via highly specific but short-lived association." In: *Trends in Biochemical Sciences* 37.2, pp. 43–48. DOI: 10.1016/j.tibs.2011.11.002.

FACHBEREICH BIOLOGIE DER JOHANNES GUTENBERG-UNIVERSITÄT MAINZ

Archiv-Nr. / 20

Zusammenfassung der Dissertation von

Dalmira Hubrich

Thema:

Systematic Interaction Interface and Variant Characterization using Protein Interaction Profiling

Eine vielversprechende Strategie, das Edgotyping, nutzt Protein-Protein-Interaktionsnetzwerke (PPI) zur systematischen Charakterisierung von Varianten und zur Aufdeckung von Krankheitsmechanismen, indem getestet wird, wie sich pathogene und nicht charakterisierte Mutationen auf Proteininteraktionen auswirken (**Vidal et al. 2011; Sahni et al. 2013; Sahniet al. 2015; Fragoza et al. 2019; Cheng et al. 2023**). Es wurden mehrere Versuche mit dem Edgotyping-Ansatz durchgeführt. So identifizierten Sahni et al. (2015) 197 Mutationen in 89 Wildtyp-Proteinen, wobei 26% einen vollständigen Verlust der Interaktion aufwiesen, 31% edgetisch waren und 43% unverändert blieben, unter Verwendung von Y2H. Unter Verwendung desselben Ansatzes bewerteten Fragoza et al. (2019) 1.676 Missense-Varianten in 4.109 Protein-Varianten-Interaktionen und identifizierten 298 störende Varianten, die 669 PPIs betreffen, und untersuchten deren funktionelle Auswirkungen. Dieser experimentelle Ansatz hat das Potenzial, uncharakterisierte Varianten zu adressieren, bleibt aber aufgrund der großen Anzahl dieser Varianten teuer und arbeitsintensiv.

In dieser Arbeit wird eine systematische Strategie zur Charakterisierung von Varianten anhand von vorhergesagten und experimentell validierten DMI-Schnittstellen durch PPI-Profiling vorgeschlagen. Kapitel 2 befasst sich mit der Erstellung der ORFeome-Sammlung, der Bewertung der BRET-Empfindlichkeit und der Entwicklung einer Klonierungspipeline mit mittlerem Durch-satz sowie eines BRET-basierten Assays zur experimentellen Validierung von DMIs.

Artikel I zeigt die erfolgreiche Anwendung von Klonierung und ortsgerichteter Mutagenese mit geringem Durchsatz, gefolgt von BRET-Assays mit mittlerem Durchsatz, um DMI-vermittelte PPIs, die am Spleißen beteiligt sind, experimentell zu validieren. Artikel II stellt eine optimierte plattenbasierte Klonierungspipeline und einen BRET-Assay zur Validierung neuartiger vorhergesagter Schnittstellen und zur Analyse von Mutantenpositionen und deren Potenzial zur Unterbrechung dieser Schnittstellen vor, basierend auf vorhergesagten Schnittstellenstrukturen aus AF-MM. Somit zeigen sowohl Artikel I als auch Artikel II die Etablierung eines systematischen Ansatzes für die experimentelle Validierung mutmaßlicher DMIs. Kapitel 3 beschreibt die Entwicklung des DMI-Prädiktors und seine Anwendung bei der Vorhersage und Kartierung dieser DMIs anhand von HuRI-PPI-Daten, gefolgt von der

Integration von ClinVar-Mutationsdaten mit vorhergesagten DMIs. Außerdem wird die Auswahl der PPIs beschrieben, die mit DMIs kartiert wurden und für eine experimentelle Validierung geeignet sind. Im Ergebnis konnten wir 46 von 96 PPIs mit dem BRET-Assay bestätigen. Von diesen haben wir die vorhergesagten DMIs von 22 Interaktionen weiter ausgewertet und 6 Kandidatenproteine (CTBP1, WWOX, PPP3CA, REPS1, SPOP und IQCB1) identifiziert. Dies zeigt, dass die Integration von DMI-Informationen mit der Erstellung von Proteininteraktionsprofilen dazu beitragen kann, Varianten zu charakterisieren, unser Verständnis der Auswirkungen von Varianten auf PPIs zu verbessern und tiefere Einblicke in die Mechanismen dieser Varianten und ihre potenziellen Beiträge zu Krankheiten zu liefern.

Genehmigt vom 1. Gutachter / von der 1. Gutachterin __

Dalmira Hubrich

+49 157 34517760 / dalmiramer@gmail.com / Mainz, Germany / LinkedIn / GitHub

Profile

As a Systems Biologist with a focus on protein networks and interfaces, I gained a solid background in systematic experimental biology during my PhD, where I also began learning computational skills, including Python and SQL. While my primary expertise lies in experimental techniques, I am now expanding into bioinformatics and computational biology, aiming to work more extensively with biological data. I also have a growing interest in artificial intelligence and its applications in biological research, with a focus on enhancing my computational skills.

Professional Experience

Researcher **December 2020-present**
Institute of Molecular Biology, Germany

- Established and adapted several techniques (e.g., cloning, site-directed mutagenesis, BRET assay, bioluminescent imaging) in the lab.
- Curated, extracted, and visualized diverse biological datasets required for my study.
- Provided experimental data and visualization support to PhD students and colleagues.
- Delivered results and contributed to collaborations across multiple projects.

Junior Data Scientist (Part-time) **April 2023 - August 2023**
Be Factory UG

- Curated, cleaned, and processed large datasets.
- Developed a tool for feature extraction and trained a machine learning model to evaluate products based on score results.
- Built entity relationships model in SQL to manage data more effectively.
- Automated data management processes to streamline workflows.

Education

- **Doctor of Philosophy in Life Sciences**|Johannes Gutenber University, Germany|December 2020 - present
- **Master in Protein Enginnering and Biochemistry**| Okinawa Institute of Sci & Tech| August 2017-2019
- **Bachelor of Biological Sciences**|Nazarbayev University| September 2011-August 2016

Skills

- **Proficient in conducting systematic experimental assays** to detect protein-protein interactions (PPIs), including BRET and ITC.
- **Experienced** with high-content screening equipment, such as Opera Phenix, for live-cell imaging and working with software like Harmony for comprehensive image analysis.
- **Trained** in Python OOP and relevant packages, including pandas, scipy, and numpy for data management and analysis; matplotlib and seaborn for data plotting and visualization; scikit-learn for machine learning models .
- **Proven track record** with over four years of experience successfully overseeing various projects and collaborating with interdisciplinary teams.
- **Competent** in presenting complex concepts to diverse audiences
- **Organized, adaptable,** and always eager to learn and deliver results.
- **Experienced** with software tools like Git, Bash, Visual Studio Code, PyCharm, SciWheel, Microsoft 365, Miro, Notion, and Adobe Illustrator.
- **Languages**: English (fluent), Russian & Kazakh (native speaker), German (A2) and ongoing